12-15-2014

Pattern Extraction From Spatial Data - Statistical and Modeling Approches

Hu Wang
University of South Carolina - Columbia

PATTERN EXTRACTION FROM SPATIAL DATA - STATISTICAL AND MODELING APPROACHES

by

Hu Wang

Bachelor of Science
China Agricultural University, 2008

Master of Science
China Agricultural University, 2010

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Geography

College of Arts and Sciences

University of South Carolina

2014

Accepted by:

Diansheng Guo, Major Professor

Sarah Battersby, Committee Member

David Hitchcock, Committee Member

Michael Hodgson, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

ACKNOWLEDGEMENTS

Throughout my four and half years of graduate study at University of South Carolina, my advisor, Dr. Diansheng Guo, guided me with his fierce intuition, breadth knowledge and spirit of practicality. His vision, passion and persistence to high standards will keep on inspiring me for many years on the road. Without his guidance and encouragement, the accomplishment of this dissertation cannot be achieved. I would also like to express my thankfulness toward my great committee, Dr. Sarah E. Battersby, Dr. David Hitchcock and Dr. Michael E. Hodgson for their mentorship and guidance. I would like to extend my thanks to all of the faculty, staff, and graduate fellows in the department of Geography at USC.

I gratefully acknowledge the funding support from the NSF CAREER (BCS-0748813) Grant and the China Scholarship Council that made my Ph.D. work possible.

I feel grateful for always having a group of friends sticking around. I enjoyed many dinner conversations with Qian, Xi, Peng and Queenie. I was fortunate to have Jiang, Ting, Xi, Chao and Yuan as wonderful officemates along the years. My friends at USC, Shufan, Emily, Hai, Junyu, Jiayu, Ke, Lei, Emma, Caglar, and many others shared invaluable experience with me along the years.

Last but not least, I would like to express my deepest gratitude to my family - Mom, Dad, Luni and Zhi - for always being there with their unreserved support and love!

ABSTRACT

Exploratory and statistical spatial data analyses are commonly used in a wide range of research fields, such as epidemiology, disease surveillance and crime analysis. Spatial epidemiology, for example, needs to detect significant spatial clusters of disease incidents to help epidemiologists identify environmental factors and spreading patterns associated with certain diseases. Existing spatial analysis approaches mostly focus on the analysis of *spatial lattice data*, i.e., observations organized by locations such as county or census tract. With the wide spread of location-aware technologies such as GPS and smart phones, *spatial interaction data* have become increasingly available, e.g., human daily mobility, traveling and migration.

The goal of this dissertation work is to develop new methodologies for the analysis of both spatial lattice data and spatial interactions data, with a focus on statistical and modeling approaches. The contribution of this dissertation includes three new methodologies for spatial scan statistics (Chapter 2), flow scan statistics (Chapter 3), and spatial interaction modeling (Chapter 4).

The first developed methodology is a new spatial scan statistic incorporating smoothing and regionalization techniques. The contribution is three-fold: 1) the new method can detect irregular shaped spatial clusters, which is more efficient and effective than existing methods; 2) the method can alleviate the multiple-testing problem by dramatically reducing the cluster search space with hierarchical regionalization; and 3)

the integration of a smoothing strategy addresses the small-area problem and significantly improves the accuracy of cluster detection. The new method is evaluated with a series of benchmark data that are widely used in related literature.

The second approach, a new flow scan statistic, is specifically designed for spatial interaction data to detect significant flow clusters. To my best knowledge, it is the first scan statistics approach for spatial interaction data that can extract significant flow clusters from very large origin-destination (OD) data sets such as migration and taxi trips. The developed flow scan statistic method scans a given OD data set with a flow tube, which is defined by a neighborhood at the origin and a neighborhood at the destination, to detect significantly higher-than-expected flow clusters among locations. The test statistic is based on the Generalized Likelihood Ratio (*GLR*), which is specifically designed to work with both area-based and point-based spatial interaction data. The new method is demonstrated and evaluated with case studies of the county-to-county migration data in U.S. and a synthetic point-based OD flow data.

The third method presented in this dissertation is a spatial interaction modeling and analysis framework that consists of (1) a piece-wise spatial interaction model to understand global flow patterns; (2) an extended spatial autocorrelation statistics based on Moran's *I* to examine the spatial distribution of model residuals; and (3) a new mapping approach to visualize local flow patterns (spatial clusters of model residuals) that cannot be explained by the configured model and global patterns. The developed model takes into account the distance, origin/destination sizes and an accessibility measure for each flow. The model outcomes (i.e., coefficients) reveal interesting global patterns, followed with the statistical analysis and mapping of model residuals, with

which one can further investigate local deviations from global trends and be able to gain a comprehensive understanding of the complex patterns hidden in spatial interaction data. A case study is carried out to analyze the migration among Metropolitan Statistical Areas (MSAs) in the United States. The major contribution of proposed framework includes a framework to configure piece-wise spatial interaction models, an extended Local Moran's *I* statistic for analyzing flow residuals, and a novel mapping method for visualizing the flow residual patterns.

The first and second approaches focus on scan statistics, with the first one improving existing spatial scan statistics by detecting irregular-shaped clusters based on regionalization and smoothing while the second approach is a new scan statistics method for analyzing spatial interaction data (i.e., location-to-location flows). The second and third methods are both for the analysis of spatial interaction data, with the former focusing on detecting significant flow clusters by developing a new statistics and the latter focusing on an exploratory framework and new approaches for spatial interaction modeling and residual analysis.

The series of new methodologies and framework introduced in this dissertation can be extended in the future to analyze spatiotemporal patterns in spatial interaction data. In this dissertation I focus on migration data analysis, while the methodologies can also be used in other many other spatial data applications, such as economic activities, trade analysis, animal migration, and disease spread.

TABLE OF CONTENTS

Acknowledgements iii

Abstract iv

List of Tables ix

List of Figures x

CHAPTER 1 : Introduction 1

CHAPTER 2 : A Spatial Scan Statistic Method with Smoothing and Regionalization 6

2.1 Abstract 6

2.2 Introduction 7

2.3 Related Work 11

2.4 Spatial Scan Statistic with smoothing 13

2.5 Evaluation and Comparison 19

2.6 Conclusion and Discussion 26

CHAPTER 3 : A Flow Scan Statistic Method for Spatial Interaction Data 31

3.1 Abstract 31

3.2 Introduction 31

3.3 Related Works 33

3.4 Methodology 38

3.5 Data and Results 43

3.6 Discussion and Future work 52

CHAPTER 4： An Exploratory Approach to Spatial Interaction Modeling and Residual Analysis 54

4.1 Abstract 54

4.2 Introduction 54

4.3 Background 57

4.4 Migration Data 68

4.5 Spatial Interaction Modeling of Migration 70

4.6 Residual Mapping and Exploratory Analysis 75

4.7 Discussion and Conclusion 84

CHAPTER 5 : Conclusion 86

5.1 Summary of Results 86

5.2 Limitations and Future Directions 89

References 90

LIST OF TABLES

Table 2.1: Cluster information. 21

Table 2.2: Comparison results for statistical power..... 28

Table 2.3: Comparison results for sensitivity.. 28

Table 2.4: Comparison results for ppv..... 28

Table 2.5: Comparison results for misclassification rate. 29

Table 2.6: Robustness analysis on kernel functions.. 29

Table 2.7: Robustness analysis on smoothing population threshold.. 30

Table 3.1: Counts of flow tubes under different criteria. 46

Table 3.2: Seven designed point-based flow clusters. 49

Table 3.3: List of all reported flow tubes.. 52

Table 4.1: A survey of empirical studies with gravity models 59

Table 4.2: Parameter estimates of the piecewise gravity model fitted to continental U.S. commuting data by 3109 counties. 61

Table 4.3: Amount of migrants between MSAs and rural areas by Age Groups. 73

Table 4.4: Calibration of migration models for seven population groups and the entire population. 74

LIST OF FIGURES

Figure 2.1: Demonstration of neighborhood definition for smoother..... 15

Figure 2.2: The simple merge algorithm..... 16

Figure 2.3: The hierarchical merge algorithm. 18

Figure 2.4: Simulated data clusters (shaded areas) for the North-eastern U.S.. 20

Figure 2.5: An illustration of accuracy measures.. 22

Figure 3.1: Illustration of community construction. 35

Figure 3.2: Flow map of migration from California from 1995 – 2000. 36

Figure 3.3: Illustration of a flow tube. 38

Figure 3.4: Top 10,000 migration flows out of 721,433 among 3075 counties for Census 2000.. 44

Figure 3.5: Significant migration flows for entire population (*p-value* < 0.001). 45

Figure 3.6: Significant migration flows without double-sided directions for entire population (*p*-value < 0.001)..... 46

Figure 3.7: Significant migration flows for age above 65 in Census 2000 (*p-value* < 0.001). 47

Figure 3.8: Significant migration flows without double-sided directions for age above 65 (*p*-value < 0.001). 47

Figure 3.9: Histograms of *LGLR* values under null hypothesis for entire migrants (left) and older migrants (right).. 48

Figure 3.10: Synthetic point-based flow data and flow scan results..... 50

Figure 3.11: Top flow tubes using flow scan statistics with multi-level nearest points.. . 51

Figure 4.1: Results in (Simini et al. 2012) and results of my gravity model..... 62

Figure 4.2: Illustration of weight matrices.. 66

Figure 4.3: Net migration ratio for 358 MSAs with Census 2000 migration data..... 68

Figure 4.4: Competing destination of 358 MSAs for entire population. 69

Figure 4.5: Distance breakpoint for each population group, determined by maximizing the Log-likelihood value in Piecewise Poisson regression. 71

Figure 4.6: Parameter values for *short-distance* migration for each age group..... 72

Figure 4.7: Parameters of *long-distance* migration for each age group..... 72

Figure 4.8: Top 1000 net residuals for entire population, overlaid by *NMR*. 76

Figure 4.9: Flow neighborhood..... 77

Figure 4.10: Spatial autocorrelation of net residuals for all migration, overlaid by *NMR*. 79

Figure 4.11: Spatial autocorrelation of net residuals for age group 5-14, overlaid by *NMR*. 81

Figure 4.12: Spatial autocorrelation of net residuals of age group 15-19, overlaid by *NMR*. 81

Figure 4.13: Spatial autocorrelation of net residuals for age group 20-24, overlaid by *NMR*. 82

Figure 4.14: Spatial autocorrelation of net residuals for age group 25-29, overlaid by *NMR*. 82

Figure 4.15: Spatial autocorrelation of net residuals for age group 30-44, overlaid by *NMR*. 83

Figure 4.16: Spatial autocorrelation of net residuals for age group 45-59, overlaid by *NMR*. 83

Figure 4.17: Spatial autocorrelation of net residuals for age group above 60, overlaid by *NMR*. 84

CHAPTER 1 : INTRODUCTION

Exploratory and statistical spatial data analyses are commonly used in a wide range of research fields. For example, in spatial epidemiology, the detection of significant disease clusters can help epidemiologists identify environmental factors and spreading patterns associated with disease and therefore direct investigation of particular disease (Aamodt et al. 2006). For such analysis, unsupervised classification (e.g. clustering) methods and statistical inference approaches are both required. The detection of statistically significant spatial clusters is a critical task in epidemiology, disease surveillance and crime analysis (Duczmal et al. 2006).

Spatial scan statistic is widely applied in detection of geographical clusters, e.g., areas with significantly high rates of disease or crime (Ceccato 2005, Heffernan et al. 2004, Kulldorff 1997). In general, the methods of spatial scan statistic all follow similar steps: 1) scan the study region with a scanning window of various sizes and limited choice of regular shapes (e.g., circle); 2) calculate a likelihood statistic for each associated scanning window; 3) consider the scanning window with the highest statistic value as the candidate cluster; 4) obtain a null distribution of the statistic through Monte Carlo simulations, and derive a *p*-value for the candidate cluster. The drawback of traditional scan statistics is that its scanning window is of certain regular shape (e.g., circle or ellipse) and consequently it might miss important clusters of different and irregular shapes.

Increasing amount of spatial interaction (SI) data has become available with technology development. SI data represents the movements over space such as human migration, daily travels, commodity flows, and information spread. Human mobility, a unique type of SI data, represents the flows of individual-moving from one location to another. It is of great importance to identify the patterns and trends of human mobility, which is useful for various domains such as epidemiology, demography, urban planning and development, tourism, transportation, and so on. From a network perspective, spatial interactions form a complex graph embedded in space, where locations are nodes and interactions are connections among locations. Nodes and connections also have attributes/characteristics.

There are two major methodology types for SI data analysis: spatial interaction modeling and exploratory network analysis. Spatial interaction models assume that the flow volume between two places is to some degree associated with the properties of the two places (e.g. population or gross flow) and the flow connection (e.g., distance) (Erlander & Stewart 1990). For example, migration flows may be correlated to various factors such as distance between two locations and employment opportunities at different locations. Spatial interaction modeling intends to estimate flow volume between each pair of origins and destinations based on a set of selected factors (Barthélemy 2011, Jung et al. 2008, Kaluza et al. 2010, Balcan et al. 2009).

The second type of methodologies for spatial interaction analysis is based on exploratory and network approaches, which aim to extract non-trivial network structures from spatial interaction data (Fortunato 2010, Guo 2009, Newman 2006, Thiemann et al. 2010). Human mobility networks are embedded in the geographic space, where network

structures (e.g., community structure or clusters) have explicit spatial meanings such as neighborhoods and functional regions. Existing methods usually focus on either network properties (with spatial variables such as distance) or the spatial distribution of network measures related to connections and nodes. There is much less attention paid on the combined complexity of both spatial distribution and network structure.

This dissertation develops three new methodologies for the analysis of spatial lattice data and spatial interaction data with a focus on statistical and modeling perspective.

In the first paper (Chapter 2), two methods of spatial scan statistics with a simple and a hierarchical merge procedure are developed for geographic cluster detection. Traditional spatial scan statistics might miss irregular clusters since their scan window of shape is of limited choices (i.e. circle or ellipse). The new methods presented in this study are based on regionalization approaches to detect spatially contiguous clusters with optimization approaches. Smoothing techniques are also integrated to obtain stable statistical measures, which can alleviate the small-area rate problem and avoid oversized clusters with extraordinary shape. Benchmark data sets with circular and irregular clusters are used to assess the new methods. Comparisons with the circular, elliptic, and double-link constraint spatial scan statistics are conducted. The proposed methods have three major contributions: 1) they are able to detect clusters with irregular shape, without defining a specific scanning window; 2) they dramatically reduce the number of cluster candidates and alleviate multiple-testing problems by building cluster candidates through regionalization; and 3) they alleviate small-area rate problem and avoid oversized clusters with extraordinary shape.

The second paper (Chapter 3) describes a new flow scan statistic method for spatial interaction data. Different from existing spatial scan statistics, which just use a scan window, the proposed Flow Scan Statistic method adopts a flow tube, which is defined by a base window on the flow origin and the other window on the destination, to scan spatial interaction data and detect significant flow clusters. The construction of flow tubes is based on the population or inflow/outflow volume at each location, which controls flow cluster size and reduces computational complexity. A flow Poisson Generalized Likelihood Ratio, which does not depend on population, is proposed to serve as a test statistic. The new approach employs Monte Carlo simulation to produce a null distribution for the test statistic. The proposed flow scan statistic can be applied to both area-based and point-based spatial interaction data, demonstrated by two case studies: 1) internal county-to-county U.S. migration data in Census 2000, and 2) a synthetic point-based data set. The evaluation results show that the Flow Scan Statistic has a good detection power, that it is not sensitive to pre-defined flow tube sizes. The uniqueness of flow scan statistic is that it not only clusters the data based on flow weights, but also determines the significance by taking advantage of Monte Carlo simulation.

The third paper (Chapter 4) presents an exploratory framework for the residual analysis of fitted spatial interaction models. The proposed framework consists of three stages: 1) fitting a spatial interaction model with the piecewise Poisson regression, taking distance, masses of locations and a competing destination variable into consideration; 2) extending the Local Moran's *I* statistic to examine the spatial distribution and clustering of model residuals; and 3) applying a new mapping approach to visualize local flow patterns (spatial clusters of model residuals) that cannot be explained by the configured

model and global patterns. The model outcome captures the global trends and the autocorrelation and mapping discovers hidden patterns that cannot be explained by the fitted model in the first stage. The framework is applied to internal U.S. migration between 358 Metropolitan Statistical Areas for seven age groups. The determined significant distance breakpoints are range from 590-1410 km by configured models. Significant clusters of flow prediction residuals are identified for seven age group migration data in terms of the proposed Flow Local Moran's *I*. The results suggest that the framework performs well for all seven age groups. The major contribution of proposed framework is to extend Local Moran's *I* to examine spatial interaction model residuals which represent the impacts of hidden factors other than the ones considered in modeling stage.

Although the three papers are separated, they are connected in several ways. Both the first and second papers are concentrating on scan statistics: the first one improves the existing spatial scan statistics by detecting irregular cluster, and the second one extends it to investigate higher-dimensional data (spatial interaction data). The second and third ones focus on better understanding spatial interaction data. The second one aims at extracting significant clusters of spatial interaction, and the third one examines the local associations of spatial interaction residuals.

CHAPTER 2 : A SPATIAL SCAN STATISTIC METHOD WITH SMOOTHING AND REGIONALIZATION

2.1 ABSTRACT

Spatial scan statistics are commonly used for detecting geographic clusters, e.g., areas with significantly excessive concentration of disease incidences or crimes. Existing methods of spatial scan statistics often adopt an exhaustive search strategy to identify clusters with regular shapes (e.g. circle or ellipse). In this chapter, I present two new methods of spatial scan statistics with smoothing and regionalization techniques, each of which (1) apply a smoothing technique to each unit to get reliable incident rates; 2) use simple or hierarchical merge strategies to aggregate data into a set of spatially contiguous regions (i.e., cluster candidates) to maximize the Likelihood Ratio; and (3) test the significance of regions (cluster candidates) with a Monte Carlo permutation. These new approaches have three main advantages over existing methods. First, they can detect significant spatial clusters of different shapes and sizes. Second, the number of candidate clusters being evaluated is much smaller, dramatically alleviating a multiple-testing problem and reduce the computational complexity. Third, the integration of smoothing technique can alleviate small-area rate problem and avoid oversized clusters with bizarre shape. I use benchmark data sets with circular and irregular clusters to evaluate the new methods and compare the results with the circular, elliptic, and

double-link constrained spatial scan statistics methods. Robustness analysis suggests that new approaches are not sensitive to the choice and setting of smoothing functions.

2.2 INTRODUCTION

Detection and evaluation of statistically significant spatial clusters is a crucial task in epidemiology, disease surveillance and crime analysis (Duczmal et al. 2006). Spatial scan statistic is commonly used for the detection of particular geographical clusters, e.g., areas with significantly high rates of disease or crime (Ceccato 2005, Heffernan et al. 2004, Kulldorff 1997). Conceptually, a spatial scan statistic method takes three steps: (1) search through all candidate clusters (e.g., areas around different locations and of different sizes and shapes) and calculate a statistical measure for each candidate cluster; (2) use a Monte Carlo permutation to generate a large number of random data sets under the null hypothesis, repeat step (1) for each random data set, and thus establish an empirical distribution of the statistical measure under the null hypothesis; and (3) assign a *p*-value (i.e., significance level) to each cluster based on its measure value (from step 1) and the null distribution (from step (2)).

However, the number of candidate clusters is often extremely large even for a moderate-sized data set, making it infeasible to enumerate all possible clusters. To alleviate this problem, existing methods often take three approaches. One is to assume a fixed shape of candidate clusters (e.g., circle, ellipse), which would dramatically reduce the number of potential clusters and make it computationally tractable (Kulldorff 1997). Such an assumption of a cluster shape, however, might cause the miss of important clusters of different and irregular shapes. For example, the widespread of disease along a river may not be of a circular shape, thus would not be captured by a circular candidate

cluster. The second kind of alleviation approach is to allow a more flexible shape definition (e.g., an ellipse (Kulldorff et al. 2006)) or incorporate a shape measure (e.g. compactness) in the statistical measure (Assunção et al. 2006, Duczmal & Assunção 2004, Duczmal et al. 2007, Duczmal et al. 2006), and then use a heuristic-based approach (such as genetic algorithms or Tabu optimization) to search clusters without enumerating all possible candidates. These approaches, however, also have their own limitations and challenges. First, they involve a number of subjective parameters (e.g., shape measure), which are difficult to configure and interpret. Second, the number of candidate clusters to be evaluated is still very large, which not only makes the search process very time-consuming but also leads to the multiple-testing problem. Moreover, candidate clusters would substantially overlap with each other and thus the tests of different candidate clusters cannot be assumed as independent, which not only adversely impacts the statistical testing power but wastes substantial computing time in evaluating unnecessary candidates as well. The third kind of approach solves this problem by adding certain screening criteria. Patil & Taillie (2004) proposed an Upper Level Set clustering detection which reduces the size of spatial cluster candidates by only considering the connected components of possible upper level sets. Tango (2008) investigated spatial scan statistic with restricted likelihood ratio which added a screening criterion in measure formula.

An alternative to reduce the computation consuming time is to avoid randomization testing. Neill et al (2006) provided a Bayesian method for Spatial Scan Statistics, which incorporated prior information and estimated the posterior probability of each cluster candidate. Chan (2009) replaced the maximum likelihood ratio with average

likelihood ratio, which allowed bypassing the Monte Carlo procedure, and they suggested that the average likelihood ratio statistic is more superior than the maximum statistic.

Assuncao et al (2006) introduced a graph structure partition method based on minimum spanning tree to control the candidate size, and they claimed that Upper Level Set method of Patil and Taillie is one particular case of theirs. However, Costa et al (2012) argued that Assunocal's method would cause the octopus effect, oversized cluster with extraordinary shape (Duczmal & Assuncao 2004), Instead, they presented an improved approach by integrating double-connected constraint to achieve a balance between likelihood maximization and cluster compactness. Although this method preserves a compact shape of cluster, adding double connected constraint in the candidate construction is too arbitrary to step across certain inconsistent unit due to the spurious data variation. It is known that the disease data for each unit may be unstable if the base population is too small, which means that the rate for small areas within a true disease cluster may reveal spurious variation. In general, the variation could be alleviated by area aggregation or choosing a higher analysis scale. In other words, double connected constraint leads that the cluster detection heavily depends on the choice of spatial scales. For example, given a cluster detected at county-level, if the same data is scaled down to a lower level (i.e. tract-level), variation of study rate at the lower level (i.e. tract) will be higher in cluster area than the one at the higher level (i.e. county). A unit with relatively low rate in the true cluster might lead to the miss of the cluster due to the double-connected constraint.

Regionalization is a spatial analysis technique that concerns the aggregation of a large number of spatial units into a small number of non-overlapping and spatially

contiguous regions while optimizing an objective function. REDCAP (Guo 2008) is a family of hierarchical regionalization methods that are based on contiguity constrained hierarchical clustering and partitioning. The REDCAP methods take two steps: (1) construct a spatially contiguous tree by enforcing a contiguity constraint in a hierarchical clustering method, e.g., the average-linkage, complete-linkage, or the Ward clustering method; and (2) partition the tree to generate a hierarchy of homogeneous regions while optimizing a within-region homogeneity measure, e.g., the sum of squared differences (SSD). REDCAP methods can also be integrated with smoothing techniques, such as empirical Bayes smoothing or kernel-based smoothing, to reduce the impact of spurious data variation due to the small-area problem and significantly improve the quality of constructed regions (Guo & Wang 2011).

In this paper, two new approaches to spatial scan statistics are presented, which neither assume a fixed shape nor evaluate a huge number of candidates. The proposed approaches incorporate smoothing technique to reduce the influence of spurious data variation, and generate cluster candidates based on simple merge or adaptive merging strategies, which are borrowed from regionalization methods. The smoothing technique is purposely brought in to overcome the limitations of random units, which could result in consistent detection ability at different scales. In addition, the integration of smoothing can help avoid the octopus effect by borrowing information from neighbors. Benchmark datasets with circular and irregular clusters from existing literature (Duczmal et al. 2006, Kulldorff et al. 2003) are used to evaluate the new methods and compare them with the well-known circular and elliptic spatial scan statistic methods in SaTScan (Kulldorff, 1997), and double-link constrained scan statistic from (Costa et al. 2012). I also execute

robustness analysis to determine the sensitivity to smoothing functions and neighborhood definitions.

In Section 2.3, I provide the necessary background of Kulldorff's and double-connected spatial scan statistic. Section 2.4 presents the new approaches to spatial scan statistics and Section 2.5 provides the evaluation results based on the synthetic data sets. I conclude with discussions in Section 2.6.

2.3 Related Work

2.3.1 Spatial scan statistic

Scan statistics was originally designed for one dimensional data analysis (e.g., Naus 1995) and then extended to two-dimensional geographical data (Kulldorff 1997, Openshaw et al. 1987, Walther 2010). Here I provide a brief introduction to the spatial scan statistics method in (Kulldorff 1997), which is implemented in SaTScan (available at www.satscan.org). SaTScan enumerates all possible circular areas of varying sizes and locations over the studied area. The purpose is to find the circular window(s) that has significantly high rates of certain observations (e.g., disease incidents). Let p be the risk within a window Z and q the risk outside the window Z in the studied area. The null hypothesis is that H_0: $p = q$, and the alternative hypothesis is H_a: $p > q$ (or $p < q$). With the Poisson model, the test statistic, likelihood ratio (λ), for a certain window Z is defined as:

$$\lambda = \frac{\left(\frac{O_Z}{P_Z}\right)^{O_Z}\left(\frac{O_W - O_Z}{P_W - P_Z}\right)^{O_W - O_Z}}{\left(\frac{O_W}{P_W}\right)^{O_W}} I\left(\frac{O_Z}{P_Z} > \frac{O_W - O_Z}{P_W - P_Z}\right) \quad (2.1)$$

where O_Z and P_Z denote the counts of observations (e.g. the number of disease cases) and

the population within Z, respectively; O_W and P_W are the counts of observations and population for the whole studied area. Likelihood ratio based on Bernoulli model was also provided in Kulldorff (1997). The likelihood ratio λ is computed for each window and the maximum value is recorded. In implementation, SaTScan places a circle over each observation location and then varies the radius of the circle to enumerate all possible circular windows. For example, suppose the data set has N spatial units (e.g., counties), the scan process may need to process as many as N^2 circles. Practically, circles covering more than 50% of the total population in the studied area are usually not considered as clusters in SaTScan.

After the calculation of the likelihood ratio λ for each circular window, a Monte Carlo simulation is used to generate a realtively large number (i.e. 999) of replications of the data set under the null hypothesis. For each replication, the maximum likelihood ratio among all the windows is obtained as explained above. With these maximum values of likelihood ratio, an empirical distribution can be constructed and a *p*-value can be assigned to each circular cluster according to its likelihood ratio. To detect low risk clusters, one can simply change the direction of the inequality sign in the indicator function in Equation 2.1. For computational consideration, $log(\lambda)$, i.e., log likelihood ratio (*LLR*), is usually used instead of the likelihood ratio λ in implementation.

Kulldorff et al. (2006) extended the circular scan statistic by adding ellipses as scanning windows. Elliptic scan statistic varies the shapes of elliptic windows by changing the ratio of the longer axis to the shorter axis (i.e. 1.5, 2, 3, 4, and 5) and the number of angels (4, 6, 9, 12, and 15) of the ellipse. It also introduces a non-compactness penalty, the formula of which is $[4s/(s+1)^2]^a$,where s is ratio of the major axis to the

minor axis of the ellipse, and *a* is a penalty parameter. The penalty is added as a factor in likelihood ratio calculation, which will favor more compact windows even if they have a marginally smaller likelihood ratio compared with those less compact ones.

2.3.2 DOUBLE-CONNECTED SCAN STATISTIC

The double-connected scan statistic (Costa et al. 2012) constructs cluster candidates by building minimum spanning trees based on graph theory, instead of applying a large number of scanning windows,. In graph theory, the contiguous geographical units in the studied area could be considered as an undirected connected graph, each edge of which represents a pair of geographically contiguous neighbors. This method builds a tree for each spatial unit in the studied area. When expanding the trees, the neighbor unit will be considered if at least two connections are found between the neighbor and the units in the current tree except for the first edge. The expansion stops whenever no increase in likelihood ratio or no satisfied neighbors. All the trees rooted in each unit will be considered as cluster candidates. The double-connected constraint, without explicit penalization, achieves good cluster compactness of clusters and alleviates the octopus effect. However, it is too conservative in overcoming random obstacles due to the small-area problem. Moreover, its definition of double connectedness is highly dependent on data resolution, where a doubly connected component can become disconnected when the data is represented at a finer resolution (i.e., smaller units). Consequently, it might only find partial clusters and fail to discover true patterns.

2.4 SPATIAL SCAN STATISTIC WITH SMOOTHING

To overcome the drawbacks aforementioned, I propose two new methods to

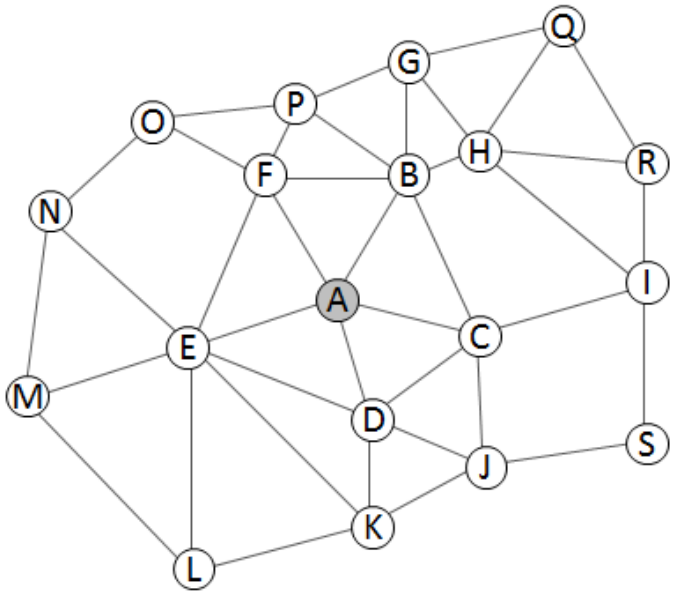
spatial scan statistic: one is based on a simple merge strategy and the other is based on a hierarchical merge strategy, both of which integrate a smoothing method. The general idea is to firstly use smoothing method to obtain a more robust measurement value for each unit to avoid the small-area problem; then construct a set of regions as cluster candidates with the smoothed values; and finally use a Monte Carlo procedure to derive a *p*-value for the test statistic for each cluster candidate. The test statistic adopted in the new methods is the log likelihood ratio (*LLR*), same as that in traditional scan statistic. There are three advantages of the new methods. First, they can detect clusters (regions) of arbitrary shapes. Second, the number of candidate clusters to be tested is very small, which helps alleviate both the multiple-testing problem and the computational burden. Third, the power of the new methods does not depend on data resolution and thus has boarder and more flexible applications.

Smoothing is a technique to alleviate the spatial variance by borrowing information from spatial neighbors besides the observation at hand. Kernel smoother is one of the most commonly used spatial smoothers. In essence, it assigns a set of weights to the neighbors of each unit in terms of kernel function. Kernel function is a distance decay function with *bandwidth*, a threshold beyond which the weight is set to zero. A kernel function can be formulated as $K_{ij}\left(\frac{d_{ij}}{h_i}\right)$, where d_{ij} is the distance between *i* and *j*, and h_i is the bandwidth. The bandwidth is typically set as the maximum distance within local neighborhood. In this study, neighborhood is defined as followed: if the sum of population within the first-order geographic neighbors is larger than the threshold, the neighborhood is set as the first-order neighbors. Otherwise, the search is extended to the next-order neighbors by distance until the threshold is satisfied. A demonstration of

neighborhood definition is shown in Figure 2.1. Threshold can be determined as a certain percentage of the total population (i.e. 1%), but the user is free to set to the most appropriate value based on their own studied problem and data. Several types of kernel functions are commonly used: Gaussian, quadratic, quartic, and etc. Smoothed cases and population size for a unit can be obtained by:

$$\hat{O}_i = \sum_j K_{ij} O_j, \text{ and } \hat{P}_i = \sum_j K_{ij} P_j.$$

These new smoothed values are used to calculate *LLR* for the associated unit. The remainder of this section introduces methodologies of two proposed scan statistics.



ID	POP	ID	POP
A	3	K	9
B	5	L	11
C	3	M	7
D	6	N	2
E	2	O	8
F	5	P	4
G	6	Q	15
H	14	R	6
I	18	S	12
J	7		

Figure 2.1: Demonstration of neighborhood definition for smoother. On the left is a graph representing contiguous information, in which two contiguous units are linked by an edge. On the right is a table listing the population size of each unit. Given that the population threshold is set to 30 here, the total population of units A's first-order neighborhood {A, B, C, D, E, F} is 24. Since it does not meet the threshold, the search is expended to the second-order neighborhood {G, H, I, J, K, L, M, N, O, P} by distance. After adding the nearest units P and H in the second-order neighborhood, the total population is increased to 42, which exceeds the threshold. Consequently the neighborhood for unit A is {A, B, C, D, E, F, P, H}.

2.4.1 SIMPLE MERGE

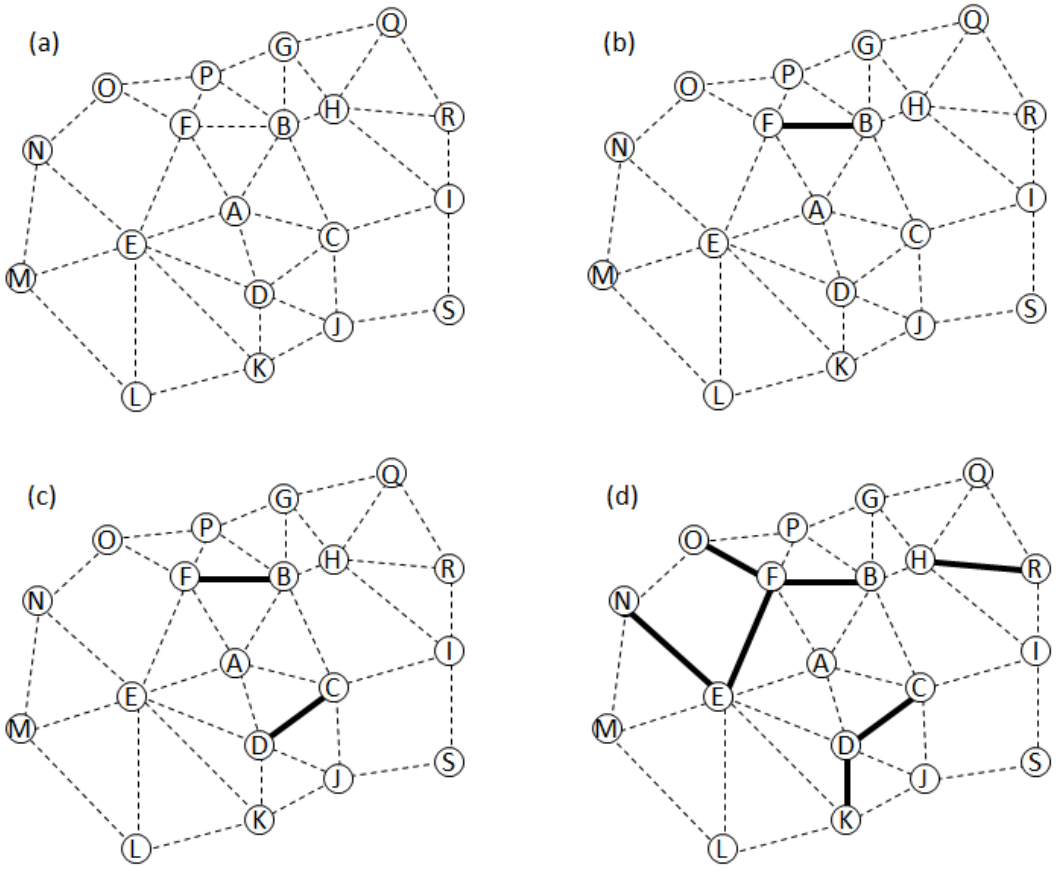


Figure 2.2: The simple merge algorithm. Initially contiguous units are linked by dotted lines as figure a. The algorithm enumerates all the links (edges) and searches the ones which provide *LLR* increases. Those links marked as solid lines are saved but not merged. Based on the marked links, all the geographically connected components are discovered and the one with the highest adjusted *LLR* value is saved as cluster candidate.

The simple-merge method starts with calculating the *LLR* value for each unit based on its smoothed values of cases and population size. Then all the neighbouring pairs in terms of the graph construction are enumerated to compare the aggregated *LLR* value with the individual *LLR* values. If the aggregated value is larger than both individual values of two nodes, then this pair would be added into a pair list. After the enumeration, geographical contiguous components are identified in the pair list. The *LLR*

values for the components are calculated based on the *adjusted* cases and population sizes. In each component, the adjusted cases and population size are defined as $\hat{O}_C = \sum_i w'_i O_i$, and $\hat{P}_C = \sum_i w'_i P_i$, where $w'_i = \max(K_{ki}), k \; is \; in \; the \; component$. The component with the highest *LLR* value is considered as the cluster candidate. The simple merge algorithm is demonstrated in Figure 2.2.

The algorithm is described as follows:

(1) For each unit *i* in the studied area,
    a. find *i*'s associated neighborhood for smoothing,
    b. apply a smoother to obtain the smoothed cases $\hat{O}_i$ and population size $\hat{P}_i$ for the unit;
    c. calculate its *LLR* value $\hat{l}_i$ based on its smoothed cases $\hat{O}_i$ and population size $\hat{P}_i$;
(2) For each pair *(i, j)* of contiguous neighbors,
    a. compute the merged *LLR* value $\hat{l}_{ij}$ based on the sums of smoothed cases ($\hat{O}_i$ and $\hat{O}_j$) and population sizes ($\hat{P}_i$ and $\hat{P}_j$);
    b. if the merged *LLR* value $\hat{l}_{ij}$ is larger than both of the individual *LLR* values $\hat{l}_i$ and $\hat{l}_j$, then put these two units into set *S*;
(3) Find all the geographically connected components in set *S*;
(4) Calculate the *LLR* values for the connected components based on the *adjusted* cases and population size;
(5) Save the component with the highest *LLR* value as the cluster candidate.

2.4.2 HIERARCHICAL MERGE

The major difference between the hierarchical-merge and the simple-merge is that the hierarchical merge method iteratively aggregates neighbours with the largest increase until all the pairs are aggregated. To aggregate two clusters in a neighbour pair, remove one of the two clusters, and update the other cluster's cases and population size as the total of two clusters' cases and population sizes, respectively. The *LLR* increases for all the pairs involving the aggregated clusters need to be updated after the aggregation. The process continues until no pair in the list. The adjusted cases and population size are

defined in the same way as the ones for geographical contiguous components in the simple merge approach. Figure 2.3 illustrates the hierarchical merge process.

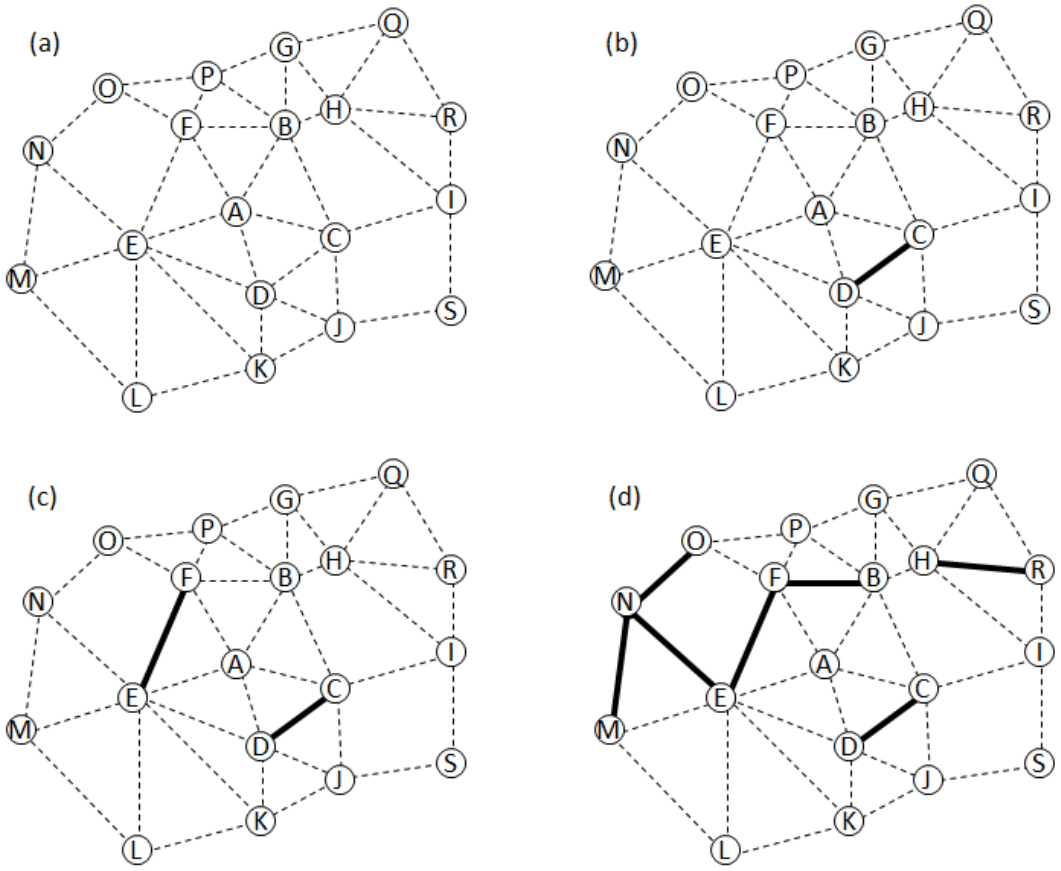


Figure 2.3: The hierarchical merge algorithm. Initially each unit is considered as a single cluster and the contiguous information is represented as dotted lines (figure a). The algorithm searches the link with largest *LLR* increase which is link (C, D) marked as solid line in figure b. Clusters C and D is merged into one cluster, and related information is updated accordingly. The process of searching largest increase is repeated until no increase for any link. The merged cluster with largest adjusted *LLR* value is picked as cluster candidate.

The steps are:

(1) For each unit *i* in the studied area,
   a. initialize the unit as a cluster,
   b. find *i*'s associated neighborhood for smoothing,
   c. apply a smoother to get the smoothed cases $\hat{O}_i$ and population size $\hat{P}_i$ for the unit;

d. calculate its *LLR* value $\hat{l}_i$ based on smoothed cases $\hat{O}_i$ and population size $\hat{P}_i$;

(2) For each pair *(i, j)* of the contiguous neighbors,
   a. compute the merged *LLR* value $\hat{l}_{ij}$ based on the sums of smoothed cases ($\hat{O}_i$ and $\hat{O}_j$) and population sizes ($\hat{P}_i$ and $\hat{P}_j$);
   b. if the merged *LLR* value $\hat{l}_{ij}$ is larger than both of the individual *LLR* values $\hat{l}_i$ and $\hat{l}_j$, then put this pair (edge) into pair list *P*;

(3) While *P* is not empty,
   a. aggregate the pair with the largest increase,
   b. update the pair list *P*. Remove one of the clusters, replace the other by the union of the two, update the increase of involved pairs;

(4) Calculate the *LLR* values for the merged clusters based on the *adjusted* cases and population size;

(5) Save the cluster with the highest *LLR* value as the cluster candidate.

2.5 Evaluation and Comparison

2.5.1 Data

To evaluate the new proposed approaches and compare the results with the circular, elliptic, and double-link spatial scan statistic methods, I use a set of benchmark data sets with three circular clusters from Kulldorff et al. (2003) and seven irregular clusters from Duczmal et al. (2006). The study region of the benchmark dataset contains 245 counties in the North-eastern United States (see Figure 2.4). The population of women from the 1990 census is used as background population, the total of which is 29,535,210. Three circular clusters are generated in rural, mixed, and urban area, respectively (Figure 2.4a); seven irregular clusters are constructed based on landscape features including Connecticut River (cluster A), Hudson River (Cluster B), Lake Ontario Coast (Cluster C), and Susquehanna River (Clusters D and E), or geopolitical boundaries of Pennsylvania (Clusters J and K) (see Figure 2.4b and c).

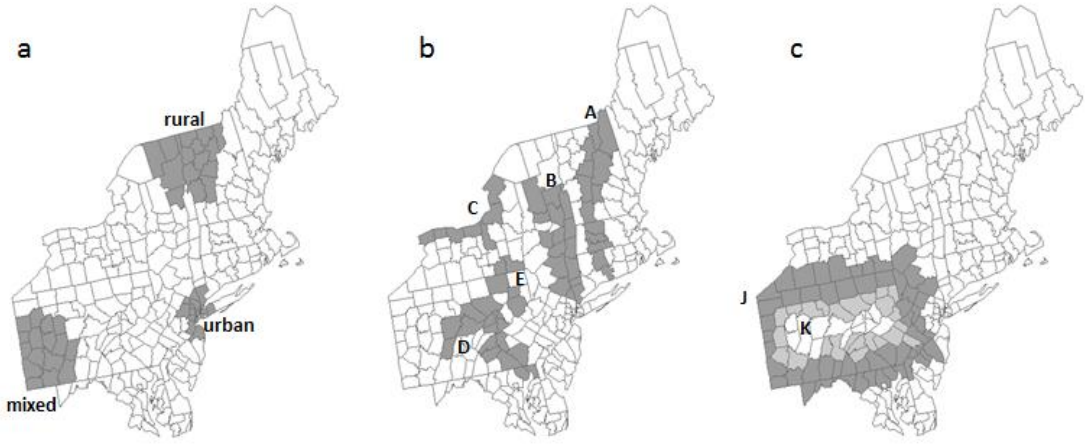


Figure 2.4: Simulated data clusters (shaded areas) for the North-eastern U.S. Circular clusters are presented in a, and irregular clusters are identified by the letters in b and c. Cluster E consists of cluster D and five nearby counties. Cluster K contains the entire Cluster J and several inner counties (lightly shaded).

There are a total of 600 simulated disease cases among the 245 counties. Outside the cluster, the cases are randomly distributed in proportion to the population under the null hypothesis. The null hypothesis (H_0) assumes that the disease rate is the same at any location and therefore the expected number of cases under H_0 is the overall disease rate times the total population in a cluster. A higher relative risk is assigned to the counties within each cluster, which is determined by the rule that the null hypothesis would be rejected with probability 0.999 while running a standard binomial test given that the true cluster locations are known. The population size, the expected number of cases under the null and alternative hypotheses, and the relative risk of each cluster are given in the Table 2.1.

As a result, there are a total of ten clusters generated (i.e. three circular and seven irregular clusters). 10000 data sets are generated to perform the comparisons for each cluster. For each data set I applied the circular, elliptic (no penalty), double-connected

scan statistic and my approaches. And 9999 random data sets are generated under the null hypothesis to estimate the cut-off point for significance. It was proved before that the results for both Poisson and Bernoulli probability models are almost identical (Costa et al. 2012), so only the results for the Poisson model are evaluated.

Table 2.1: Cluster information. $E[c|H_0]$ and $E[c|H_a]$ are the expected numbers of cases under the null and alternative hypotheses, respectively. Total number of cases is 600 (simulated).

Cluster	**Region**	**#counties**	**Population**	$E(c\|H_0)$	$E(c\|H_A)$	***Relative Risk***
Circular	Rural	16	360275	7.32	27.57	3.90
	Mixed	16	1684327	34.22	67.61	2.10
	Urban	16	7627173	154.94	208.52	1.53
Irregular	A	13	1057407	21.48	47.59	2.32
	B	16	1672387	33.97	63.44	1.97
	C	7	709519	14.41	37.52	2.71
	D	15	119235	2.42	5.52	2.29
	E	21	1483995	30.15	58.96	2.06
	J	55	3198049	64.97	99.13	1.63
	K	78	7775129	157.95	194.26	1.34

To compare with other methods, quartic kernel function with 200,000 as the population threshold is applied to my methods. Robustness analysis is also conducted by applying different kernel functions and population thresholds. The results in section 2.5.4 demonstrate that the proposed methods are not sensitive to either of these settings.

2.5.2 Evaluation Measures

Following Costa et al. (Costa et al. 2012), the two proposed approaches and other three methods are examined with respect to statistical power, and three accuracy measures: the sensitivity, the positive predictive value (*PPV*) and misclassification rate.

The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is actually false (and should be rejected) at the significance level *alpha* = 0.05. In other words, the power represents the capability to detect a cluster when it really exists. The average powers of a method with the 10,000 benchmark data sets are used for comparison.

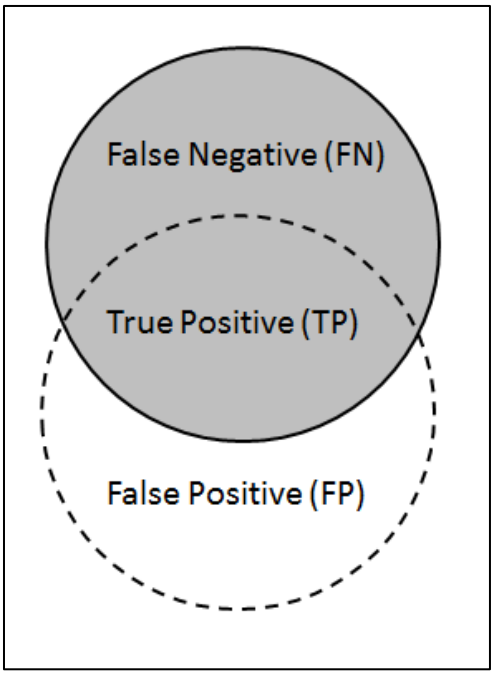


Figure 2.5: An illustration of accuracy measures. The solid circle is actual positive area, and the dash circle is detected positive area. Therefore True Positive (TP) is overlap of these two circles. False Positive (FP) is the non-shaded area within dash circle. False Negative (FN) is the proportion of solid circle falling outside the dash circle.

Apart from the power test, three accuracy measures of detection performance are used to evaluate the methods. These measures not only consider the power of a method, but also calculate the accuracy of the detection results. The measures are based on several definitions (Figure 2.5). True Positive (*TP*) is the population size of counties which is correctly detected as significant, False Positive (*FP*) is the population size of counties

which is incorrectly detected as significant, and False Negative (*FN*) is the population size of counties which is incorrectly detected as not significant.

The sensitivity measure is defined as the ratio of *TP* divided by the actual positive, which means the proportion of actually positive population who are correctly assigned to the positive group. It can be represented as $sensitivity = TP/(TP+FN)$. The value of sensitivity is ranged from 0 to 1. The closer the value is to one, the more likely an actually negative unit is to be tested as negative. Put another way, a high sensitivity value suggests that it is nearly certain that the detected negative units are actually negative (not in the high-risk area). The method with high sensitivity helps rule out high risk disease area.

The *PPV* measure is defined as the ratio of *TP* to the detected positive, which means the proportion of detected positive population who are actually positive. It can be represented as $PPV = TP/(TP+FP)$. The value of *PPV* is ranged from 0 to 1. The closer the value is to one, the more likely an actually positive unit is to be determined as positive. In other terms, a high *PPV* value indicates that it is very confident that the detected positive units are actually positive (in the high-risk area). The method with high PPV helps rule in high risk disease area.

The misclassification rate is the percentage of sum of *FN* and *FP* divided by the total population, i.e., the percentage of population who are allocated to the incorrect group. A method with small misclassification rate means that the detected cluster fits the true cluster well.

2.5.3 COMPARISON RESULTS

Table 2.2 presents the results of statistical power test for circular, elliptic, double-link, and two new proposed scan statistics with quartic kernel function. The results suggest that circular scan statistic has the best *power* for circular clusters, and elliptic scan statistic achieves the best *power* for irregular clusters. Averages of *power* results for ten different cluster types are reported at the last column. It shows that in general elliptic scan statistic has the best *power*. Simple merge method has relatively lower *power*, while hierarchical merge method has comparable power with double-link.

Table 2.3 shows the *sensitivity* results, which are used to evaluate that on average how many actually high-risk units are detected. The results indicate that both of simple and hierarchical merge methods achieve better performances than circular, elliptic and double-link scan statistic in this perspective. Double-link scan statistic gives the worst results because of its arbitrary double connected constraint.

Table 2.4 gives the *PPV* results, which represent the performance on how many detected units are actually within the high-risk areas. The results exhibit that double-link has the best performance, and hierarchical merge method gives the similar performance as circular and elliptic scan statistic. The simple merge method performs as well as hierarchical merge method in this standpoint because they wrongly include about 8% of the total population in the detected clusters on average (see Table 2.5).

Table 2.5 shows the *misclassification* ratio results, which are the percentage of the population wrongly identified in the total population. In this view, hierarchical merge method gives the best results, and simple merge method performs worse than elliptic and

double-link scan statistics. As expected circular scan statistic shows competitive results for the circular cluster models, but it performs not well for the irregular clusters.

In general, circular and elliptic scan statistics have good power results, and balanced performance in *sensitivity*, *ppv* and *misclassification* results. Double-link scan statistic performs well in *power* and *ppv*, but it gives unsatisfactory result in *sensitivity* measure. Simple merge method achieves good results in *sensitivity*, but low in *ppv*. Hierarchical merge method has comparable *power*, balanced *sensitivity* and *ppv*, and the best *misclassification* results.

2.5.4 ROBUSTNESS ANALYSIS

To answer the question that how sensitive do the different kernel functions and population thresholds effect on the output of simple and hierarchical merge methods, several settings are set up to perform robustness analysis. I examine three kernel functions, quartic, quadratic and Gaussian, and four population thresholds ranging from 0.5% to 1% of total population, 150,000, 200,0000, 250,000 and 300,000. When testing robustness on kernel functions, the size of 200,000 people is set as population threshold. Correspondingly, quartic kernel is used when testing robustness on population threshold settings. Since simple and hierarchical merge methods produce similar robustness results, here only the result of hierarchical merge is presented.

Table 2.6 and Table 2.7 show the results of measures including statistical *power*, *sensitivity*, *ppv* and *misclassification* rate on kernel functions and population thresholds, respectively. The results suggest that performances vary little with different kernel functions or population thresholds, which means that the methods are robust in these settings.

2.6 Conclusion and Discussion

This paper presents two alternative approaches to spatial scan statistics by integrating with smoothing techniques and two regionalization strategies (simple merge or hierarchical merge). Through the integration of smoothing and regionalization, unstable small-area rates can be enhanced by borrowing information from neighbours, which subsequently improves the construction of cluster candidates for statistical testing. The two new approaches have three main advantages over existing methods. First, it can detect clusters of arbitrary shapes. Second, the number of candidate clusters being evaluated is much smaller than that of the existing methods, which dramatically alleviates the multiple-testing problem. Last but not least, it can detect clusters and overcome the small-area problem, which enables the detection of clusters at different spatial scales.

Benchmark data sets, including data with circular clusters (Kulldorff et al. 2003) and irregular-shaped clusters (Duczmal et al. 2006), have been used to evaluate and compare the new methods with existing spatial scan statistics methods, including the circular, elliptic, and double-link constrained spatial scan statistics. The comparison results show that among the five methods (including those two introduced here), circular and elliptic scan statistics have the best performance on power but suffer severely from the multiple-testing problem (thus they can only report a single cluster as significant if any) and can only detect regular-shaped clusters.

To detect irregular-shaped clusters, the hierarchical merge approach of mine achieves similar performance as the existing double-link method and is better than the latter in terms of being independent of unit size and spatial resolution. Moreover, the hierarchical merge approach has the lowest *misclassification* rate, most balanced

performance on *sensitivity* and *ppv*, and in the meantime maintains an acceptable computational complexity. On the other hand, the simple merge method of mine has the best performance on the *sensitivity* accuracy measure and the lowest computational cost. Therefore, the simple merge method is recommended for detecting clusters of rare events and for applications that demand frequent and quick analysis, such as deadly disease control and severe crime hotspot analysis. Otherwise, the hierarchical merge method is recommended. Both of these two methods are not sensitive to different choice of kernel functions and neighborhood definitions.

Table 2.2: Comparison results for statistical power. For simple and hierarchical merge methods, quartic kernel function with 200,000 as population threshold is applied for comparison.

	Mixed16	**Rural16**	**Urban16**	**A**	**B**	**C**	**D**	**E**	**J**	**K**	**AVG**
Circular	**0.9492**	**0.9695**	**0.9266**	0.8530	0.7878	0.8808	0.8605	0.8068	0.6875	0.7978	0.8520
Elliptic	0.9337	0.9626	0.9136	**0.8998**	**0.8360**	**0.9087**	**0.9025**	**0.8483**	**0.7260**	**0.8242**	**0.8755**
Double-link	0.9166	0.9198	0.8120	0.8148	0.8002	0.7871	0.8922	0.7772	0.5940	0.5259	0.7840
Simple	0.9030	0.7537	0.7784	0.6953	0.6411	0.6791	0.7929	0.7712	0.6312	0.6938	0.7340
Hierarchy	0.9324	0.9170	0.7962	0.7595	0.7121	0.7535	0.8343	0.8003	0.6489	0.6883	0.7843

Table 2.3: Comparison results for sensitivity. For simple and hierarchical merge methods, quartic kernel function with 200,000 as population threshold is applied for comparison.

	Mixed16	**Rural16**	**Urban16**	**A**	**B**	**C**	**D**	**E**	**J**	**K**	**AVG**
Circular	0.8866	0.8629	**0.8906**	0.6841	0.5772	0.7322	0.6704	0.6164	0.5524	0.6591	0.7132
Elliptic	0.8513	0.7921	0.8700	0.8232	0.6789	**0.8301**	0.7119	0.5876	0.5257	**0.6416**	0.7312
Double-link	0.7530	0.5897	0.6300	0.5786	0.4884	0.5628	0.5883	0.3856	0.2604	0.1753	0.5012
Simple	**0.9184**	**0.9567**	0.8349	**0.9004**	**0.8942**	0.8141	**0.8393**	**0.7661**	**0.6139**	0.5181	**0.8056**
Hierarchy	0.8816	0.8937	0.8159	0.8654	0.7960	0.7429	0.7264	0.6208	0.5562	0.4239	0.7323

Table 2.4: Comparison results for ppv. For simple and hierarchical merge methods, quartic kernel function with 200,000 as population threshold is applied for comparison.

	Mixed16	**Rural16**	**Urban16**	**A**	**B**	**C**	**D**	**E**	**J**	**K**	**AVG**
Circular	0.8996	**0.8991**	0.8886	0.7275	0.6529	0.7323	0.5829	0.5402	0.6243	0.7067	0.7254
Elliptic	0.8199	0.8389	0.8204	**0.7681**	0.6858	0.7553	0.6827	0.6037	0.6684	0.7598	0.7403
Double-link	**0.9591**	0.8954	**0.9430**	0.7641	**0.8130**	**0.8395**	**0.8230**	**0.7543**	**0.8208**	**0.8697**	**0.8482**
Simple	0.8025	0.4563	0.8776	0.4378	0.4557	0.5174	0.4340	0.4291	0.6593	0.6932	0.5763
Hierarchy	0.8848	0.8281	0.8802	0.6199	0.6409	0.7019	0.6474	0.6220	0.6893	0.7293	0.7244

Table 2.5: Comparison results for misclassification rate. For simple and hierarchical merge methods, quartic kernel function with 200,000 as population threshold is applied for comparison.

	Mixed16	**Rural16**	**Urban16**	**A**	**B**	**C**	**D**	**E**	**J**	**K**	**AVG**
Circular	1.59	**0.26**	**6.41**	3.35	6.59	2.33	4.61	6.51	17.80	18.10	6.76
Elliptic	2.62	0.70	9.26	2.42	5.32	1.80	3.54	5.67	16.12	**15.94**	6.34
Double-link	1.55	0.65	10.05	**2.43**	**3.76**	**1.41**	**2.19**	**3.78**	14.97	22.36	6.32
Simple	2.26	3.31	7.54	5.99	7.71	3.45	6.67	8.04	13.69	17.96	7.66
Hierarchy	**1.48**	**0.54**	7.85	2.90	4.27	1.74	3.27	4.61	**13.58**	18.83	**5.91**

Table 2.6: Robustness analysis on kernel functions. Hierarchical merge method with 200,000 as population threshold is used for analysis.

	Kernel Function	**Mixed16**	**Rural16**	**Urban16**	**A**	**B**	**C**	**D**	**E**	**J**	**K**	**AVG**
Power	*quartic*	0.9324	0.9170	0.7962	0.7595	0.7121	0.7535	0.8343	0.8003	0.6489	0.6883	0.7843
	quadratic	0.9350	0.9142	0.8377	0.7261	0.6966	0.8068	0.8318	0.7928	0.6579	0.7293	0.7928
	Gaussian	0.9207	0.9113	0.7822	0.7274	0.7059	0.7651	0.8477	0.8180	0.6555	0.7144	0.7848
Sensitivity	*quartic*	0.8480	0.8772	0.7661	0.8294	0.7352	0.7077	0.5867	0.4672	0.4439	0.2876	0.6549
	quadratic	0.9029	0.8943	0.8102	0.8507	0.7968	0.7348	0.6667	0.5549	0.5026	0.3741	0.7088
	Gaussian	0.8986	0.9062	0.7839	0.8571	0.7761	0.8088	0.6823	0.5503	0.4723	0.3270	0.7063
PPV	*quartic*	0.9101	0.8359	0.9104	0.6805	0.7049	0.7875	0.7241	0.7017	0.7652	0.7689	0.7789
	quadratic	0.8901	0.8247	0.8895	0.6313	0.6415	0.7308	0.6654	0.6427	0.7290	0.7582	0.7403
	Gaussian	0.9091	0.8583	0.9092	0.6550	0.7170	0.7175	0.6960	0.6692	0.7485	0.7543	0.7634
Misclassi-fication Ratio	*quartic*	1.48	0.54	7.85	2.90	4.27	1.74	3.27	4.61	13.58	18.83	5.91
	quadratic	1.54	0.65	7.81	3.69	5.35	1.93	4.10	5.64	14.13	17.71	6.25
	Gaussian	1.36	0.55	7.47	3.57	4.64	1.81	3.51	5.06	13.56	18.16	5.97

Table 2.7: Robustness analysis on smoothing population threshold. Hierarchical merge method with quartic kernel function is used for analysis.

	Threshold	**Mixed16**	**Rural16**	**Urban16**	**A**	**B**	**C**	**D**	**E**	**J**	**K**	**AVG**
Power	*150,000*	0.9363	0.9274	0.7943	0.7485	0.7093	0.7534	0.8329	0.7938	0.6493	0.6820	0.7827
	200,000	0.9324	0.9170	0.7962	0.7595	0.7121	0.7535	0.8343	0.8003	0.6489	0.6883	0.7843
	250,000	0.9317	0.8975	0.8004	0.7639	0.7147	0.7590	0.8360	0.8027	0.6507	0.6937	0.7850
	300,000	0.9304	0.8840	0.8039	0.7699	0.7169	0.7630	0.8432	0.8056	0.6546	0.6990	0.7871
Sensitivity	*150,000*	0.8776	0.8576	0.8162	0.8562	0.7956	0.7408	0.7182	0.6151	0.5528	0.4192	0.7249
	200,000	0.8816	0.8937	0.8159	0.8654	0.7960	0.7429	0.7264	0.6208	0.5562	0.4239	0.7323
	250,000	0.8857	0.9087	0.8150	0.8680	0.8014	0.7445	0.7280	0.6234	0.5609	0.4285	0.7364
	300,000	0.8971	0.9217	0.8146	0.8714	0.8022	0.7442	0.7381	0.6360	0.5628	0.4348	0.7423
PPV	*150,000*	0.8927	0.8290	0.8802	0.6241	0.6422	0.7018	0.6531	0.6283	0.6899	0.7287	0.7270
	200,000	0.8848	0.8281	0.8802	0.6199	0.6409	0.7019	0.6474	0.6220	0.6893	0.7293	0.7244
	250,000	0.8831	0.8086	0.8803	0.6155	0.6324	0.7000	0.6443	0.6166	0.6874	0.7273	0.7195
	300,000	0.8810	0.7944	0.8804	0.6124	0.6308	0.6980	0.6371	0.6113	0.6854	0.7255	0.7156
Misclassi-fication Ratio	*150,000*	1.43	0.58	7.84	2.87	4.25	1.73	3.24	4.56	13.60	18.93	5.90
	200,000	1.48	0.54	7.85	2.90	4.27	1.74	3.27	4.61	13.58	18.83	5.91
	250,000	1.48	0.58	7.87	2.96	4.37	1.76	3.31	4.67	13.59	18.80	5.94
	300,000	1.45	0.62	7.88	3.00	4.43	1.79	3.37	4.71	13.60	18.72	5.96

CHAPTER 3 : A FLOW SCAN STATISTIC FOR SPATIAL INTERACTION DATA

3.1 ABSTRACT

The study of spatial interactions (SI), for example, human movement from one place to another, have a fundamental role in many models and studies, such as urban planning and spread of epidemics. However, the analysis of spatial interaction data remains a crucial challenge due to its complexity. This paper presents a Flow Scan Statistic for spatial interaction data analysis, which aims at revealing hidden flow patterns and making corresponding statistical inference. The proposed Flow Scan Statistic applies a large number of flow tubes to scan spatial interaction data and employs a Monte Carlo simulation procedure to generate the null distribution of the test statistic. The Flow Scan Statistic naturally works for both area-based and point-based SI data, demonstrated by two cases studies with the U.S. internal county-to-county migration data in Census 2000 and a synthetic point-based flow data for evaluation.

3.2 INTRODUCTION

Spatial Interaction (SI) data represents the movements of people, products, services, information, or any other type of flows among places. Examples of spatial interaction include migration, disease spread, travel, and trade. Understanding such location-to-location movements is critical for a wide range of researches and application domains, such as business decision making (Chun et al. 2012, Tobler 1981), urban

planning (Clark 1967), emergence management (Eubank et al. 2004, Ferguson et al. 2006, Ferguson et al. 2005, Germann et al. 2006, Guo 2007), human migration (Perry 2006, Ambinakudige & Parisi 2010, Guo 2009, Johnson et al. 2005, Brockmann et al. 2006) and resource management (McCool & Kruger 2003, Njock & Westlund 2010).

Human mobility networks, as a unique category of SI data, describes the flow of individuals moving from one location to another in a city or across a country. Typical data of human mobility includes migration, commuting, and travelling. Comprehensive analysis of SI data requires explicit consideration of its network features. The integration of geographical and network features is the most unique aspect of SI data analysis.

There are two major types of methodologies in SI data analysis. The first type is spatial interaction modeling. It is known that the flow volume between two places is to some degree related to the masses of two places (e.g. population or gross flow). Taking migration as an example, one thousand migrants between two less-populated cities have more significance than the same amount of flows between two large metropolitan areas. Additionally, migration flows may also be related to various factors such as distance and employment opportunities. One common approach to perform spatial interaction data analysis is through modeling, which aims at predicting or estimating flows between pairs of origins and destinations based on a set of selected factors.

The second type of spatial interaction analysis is through exploratory and network-based analyses, which aim at extracting non-trivial network structures from spatial interactions. Human mobility networks embedded in geographic space and network structures (e.g., community structure or clusters) have explicit spatial meanings such as neighborhoods and functional regions. Existing methods in this category usually

either focus on the network properties with spatial variables (such as distance) or the spatial distribution of connections and nodes. Much less attention has been paid to the complexity that combines both the spatial distribution and network structure. In this project, a new approach that takes into consideration the properties of geographic units and connected flows will be developed to detect significant patterns of SI data. In spirit of the scan statistic methods, the proposed approach generates a number of circle-pairs/region-pairs -- one of which represents the origins and the other represents the destinations -- to scan the entire study flows, and employ Monte Carlo simulation to identify the significant flow pairs in the data.

In Section 3.3, I present the related works including necessary background of spatial scan statistics. Section 3.4 provides the new proposed scan statistic for spatial interaction data and in section 3.5 I will apply the new approach to investigate the migration flow patterns of Census 2000, and to evaluate it based on a synthetic point-based flow data. I conclude with discussions in section 3.6.

3.3 Related Works

Existing spatial network data, especially for the human mobility, are usually large and complex. Current exploratory approaches (as opposed to the modeling approaches introduced above) for analyzing such data could be divided into two categories based on their application purposes: community detection and flow clustering. Most of existing studies of flow clustering methods are primarily based on visualization and mapping approaches. In this section, I will begin with spatial scan statistic for traditional lattice data or point data, both of which are not flow-based data. Then approaches for

community detection and flow visualization will be reviewed. Finally, several other alternative approaches will be discussed.

3.3.1 SPATIAL SCAN STATISTIC

Scan statistic was originally used for one dimensional data analysis (e.g., Naus 1995) and then extended to two-dimensional spatial data (Kulldorff 1997, Openshaw et al. 1987, Walther 2010). Spatial scan statistics (Kulldorff 1997) such as the methods implemented in SaTScan (available at www.satscan.org) are commonly used for detecting spatial clusters and searching for unusual places such as high-disease-rate regions. Although these methods are currently only applicable to location-based geographical data (e.g., lattice data or point data), the general idea is valuable and may be extended to analyze SI data.

SaTScan enumerates all possible circular areas of varying sizes and locations over the studied area. The purpose is to find the circular window(s) that has significantly high rates of certain observations (e.g., disease incidents). Let p be the risk within a window Z and q be the risk outside the window Z in the studied area. The null hypothesis is that H_0: $p = q$, and the alternative hypothesis is Ha: $p > q$ (or $p < q$). With the Poisson model, the test statistic for a certain window Z, likelihood ratio (λ), is defined as:

$$\lambda = \frac{\left(\frac{O_Z}{P_Z}\right)^{O_Z}\left(\frac{O_W - O_Z}{P_W - P_Z}\right)^{O_W - O_Z}}{\left(\frac{O_W}{P_W}\right)^{O_W}} I\left(\frac{O_Z}{P_Z} > \frac{O_W - O_Z}{P_W - P_Z}\right) \quad (3.1)$$

where O_Z and P_Z denote the counts of observations (e.g. the number of disease cases) and the population within Z, respectively; O_W and P_W are the counts of observations and population for the whole studied area, respectively. The likelihood ratio λ is computed for each window and the maximum value is recorded. In implementation, SaTScan places a

circle over each observation location and then varies the radius of the circle to enumerate all possible circular windows. For instance, suppose the data set has N spatial units (e.g., counties), the scan process may process as many as N^2 circles. After calculation of likelihood ratio λ for each circular window, a Monte Carlo simulation is employed for inference.

3.3.2 Community Structure Detection

Community detection is to characterize spatial network structures by partitioning the network into natural components (or communities). A community is considered as a set of units that have more interactions within them than in the outside units of the network (Figure 3.1). Various approaches have been proposed. Please see Fortunato (2010) for a detailed review. In particular, modularity is a commonly used measure to quantify the community structure. Many researches employed clustering methods or graph partitioning methods to optimize modularity (Guo 2009, Newman 2006, Thiemann et al. 2010). Wang et al. (2008) presented a spatial scan statistic with Poisson discrepancy for graph clustering which could be applied to detect community.

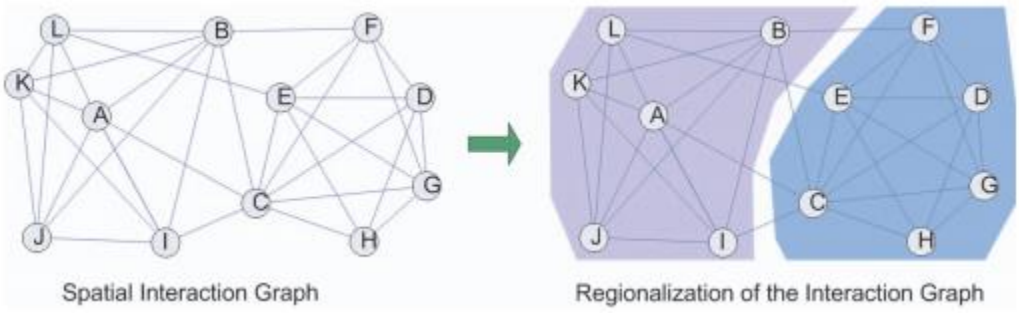


Figure 3.1: Illustration of community construction. (Source: Guo 2009)

3.3.3 VISUALIZATION APPROACHES

Visualization and flow mapping can help explore and understand SI data, similar to that graph drawing can help non-spatial graphs. In order to visualize SI data, one traditional and straightforward approach is flow map (Tobler 1977, Tobler 1981, Kwan 2000, Phan et al. 2005). The general idea of a flow map is to employ directed straight or curved lines to link the origins and destinations, and to adopt line width or color to present the flow amounts. Because of its capability of exhibiting geographical and network perspectives of SI data simultaneously, flow map has been applied and studies by many researchers (Kwan 2000, Phan et al. 2005). However, flow map has the limited capability to present SI data from medium and large data sets.

Figure 3.2: Flow map of migration from California from 1995 – 2000 (Source: Verbeek et al. 2011). Top-left is the basic flow map (Tobler 2004); top-middle is the result by bundling with crossing (Phan et al. 2005a); top-right is the map by spirals (Verbeek et al. 2011); bottom is the bundled complete migration graph by Cui et al. (2008).

To increase the capability of flow map, many researchers attempted to reduce the edges by smoothly bundling the flows (edges) (Phan et al. 2005, Cui et al. 2008, Verbeek

et al. 2011). Phan et al. (2005) proposed a method, inspired by graph layout algorithms, to produce flow maps by minimizing edge crossings while maintaining the relative position of nodes. Cui et al. (2008) adopted a control mesh to guide the edge-clustering process which can group edges into bundles and reduce the overall edge crossings. Verbeek et al. (2011) created flow map by using logarithmic spirals which naturally induced a clustering on the targets and avoided obstacles. This kind of approaches (see Figure 3.2) could show the major flow patterns and deal with the issues of edge crossings in graphical view, but ignore the flow properties.

3.3.4 Alternative Approaches

Graph theory has traditionally been employed in network analysis. However, for highly connected networks such as the U.S. migration at state level, it is not meaningful to examine its degree distribution, which is a commonly used measure of network properties. Similarly, clustering coefficient and other path-based measures are also not appropriate for completely connected networks. Donges et al. (2012) treated a network as a discrete sub-network of a "continuous" graph, which allows the use of classical statistical measures.

Recently, many researches of SI data analysis focus on data mining approaches. Laube, Imfeld and Weibel (2005) used a geographic knowledge discovery approach to discover flow patterns of point objects by comparing the flow attributes over space and across time. Gennady, Natalia and Stefan (2007) applied multidisciplinary approach to extract significant places and flows from large amounts of SI data. Fang et al. (2012) presented a spatiotemporal analysis of critical taxis trajectories based on space-time

prism concepts. Guo & Zhu (Guo & Zhu 2014) proposed a flow density estimation method to smooth the flow data and demonstrate their approach with U.S. migration data.

3.4 METHODOLOGY

3.4.1 FLOW SPATIAL SCAN STATISTIC

Instead of applying circles in spatial scan statistic, flow scan statistic uses a large number of overlapped flow tubes to define the scan window, each of which represents a possible candidate for a hot flow tube with spatial interaction inside. The flow tube consists of two regions, which represent the origin and destination of the flow, respectively (Figure 3.3). Theoretically, the flow tubes could vary with different sizes, shapes and locations for both origin and destination regions, but in practice I choose circular regions covering a pre-defined size of population (or *movers*). The details will be discussed in the next session.

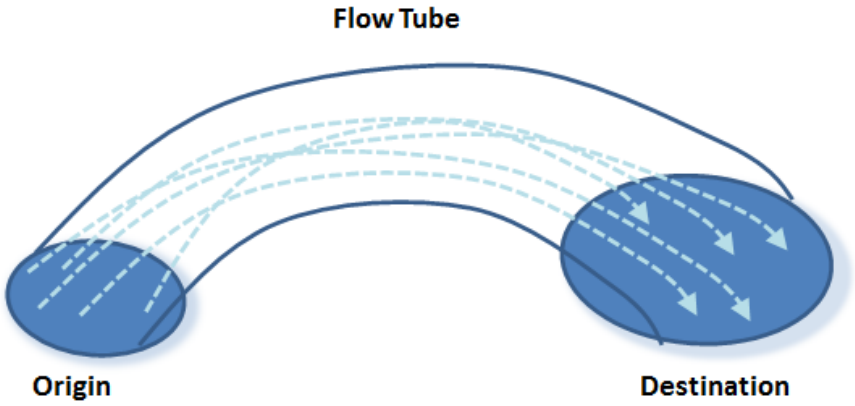


Figure 3.3: Illustration of a flow tube.

Since in most cases it is hard to obtain the data of population who have intention to move, the expected volume must be calculated using only the observed migration flow

data. Even if population data is available, it is not recommended to use population of residents as the population-to-move data because not all the residents in the area intend to move. Using population of residents would violate inference assumption and lead to biased results.

Given that the observed flow volume from area *o* to area *d* isf_{od}, the total volume of observed flows *F* is:

$$F = \sum_o \sum_d f_{od} \tag{3.2}$$

The expected value of flow $\hat{f}_{od}$ from county *o* to county *d* is calculated conditioned on the observed outflow and inflow marginal:

$$\hat{f}_{od} = \frac{1}{F} \times f_{o*} \times f_{*d} \tag{3.3}$$

Where the outflow and inflow marginal is $f_{o*} = \sum_j f_{oj}$, and $f_{*d} = \sum_i f_{id}$, respectively.

It is the proportion of all movers that are moving from area *o* times the total amount of movers that are moving to area *d*. The underlying assumption of this definition is that the probability of a mover being from area *o*, given that it is observed to be moving to area *d*, is the same for all areas. Although practically a mover from county *o* is not allowed to move to the same county *o* which might violate the assumption, the assumption could approximately hold true because $\hat{f}_{oo}$ is too small comparing to *F*. Expected volume of flows through flow tube from region *A* to region *B* FT_{AB} is the summation of these expectations from all the counties in region *A* to all the counties in region *B*:

$$\hat{f}_{AB} = \sum_{o \in A} \sum_{d \in B} \hat{f}_{od} \tag{3.4}$$

Let f_{AB} describe the amount of flows through FT_{AB}. Conditioned on the marginal, and assuming that there is no spatial interaction between flow origins and destinations

(null hypothesis), f_{AB} is distributed according to the hypergeometric distribution with mean as $\hat{f}_{AB}$ and probability function as

$$P(f_{AB}) = \frac{\binom{f_{A*}}{f_{AB}}\binom{F-f_{A*}}{f_{*B}-f_{AB}}}{\binom{F}{f_{*B}}} = \frac{\binom{f_{*B}}{f_{AB}}\binom{F-f_{*B}}{f_{A*}-f_{AB}}}{\binom{F}{f_{A*}}} \quad (3.5)$$

The function could be interpreted in two ways, represented by the two equal signs respectively. The first view is that given a certain number of people f_{*B} moving to *B*, the first equation means the probability of choosing f_{AB} out of f_{A*} (people moving from *A*) and $f_{*B} - f_{AB}$ out of $F - f_{A*}$ (people moving from other areas). The second way is from the point choosing people moving from *A*. The second equation means the probability of choosing f_{AB} out of f_{*B} (people moving to *B*) and $f_{A*} - f_{AB}$ out of $F - f_{*B}$ (people moving to other areas).

When f_{A*} and f_{*B} are small enough compared to *F*, f_{AB} is approximately Poisson distributed with mean of $\hat{f}_{AB}$. Based on this approximation, flow Poisson Generalized Likelihood Ratio (*GLR*) is applied as a measure of the evidence that flow tunnel FT_{AB} contains spatial interaction:

$$GLR = \left(\frac{f_{AB}}{\hat{f}_{AB}}\right)^{f_{AB}}\left(\frac{F-f_{AB}}{F-\hat{f}_{AB}}\right)^{F-f_{AB}} \quad (3.6)$$

In other words, *GLR* is the observed value divided by the expected to the power of the observed inside the flow tube FT_{AB}, multiplied by the observed divided by the expected to the power of the observed outside the flow tube FT_{AB}.

In order to test the null hypothesis, Monte Carlo simulation is performed in generating the distribution of *GLR* under the null hypothesis. Since population-intend-to move data is not available, the migration data cannot be permuted in the usual way for spatial scan statistic. Instead, the destinations of all the migrants are shuffled and

randomly assigned to the original origins. In this way, it is guaranteed that both of the inflow and outflow marginal are the same for each area. The permutation process is executed for 999 times, and the maximum *GLR* is calculated and recorded for each permutation. These 999 maximum *GLR* values construct a distribution of null hypothesis. According to the null distribution, *p*-value, $p = R/(L+1)$ where *R* is the rank in the maximum *GLR* list and *L* is the length of the list, is assigned to each candidate for real data. The candidates with *p*-value smaller than the significant level ($\alpha = 0.01$ or 0.05) are identified as significant hot flows in the flow map. Logarithm of *GLR* (*LGLR*) is often calculated instead to simplify the computation.

3.4.2 IMPLEMENTATION

3.4.2.1 FLOW SHUFFLE

When generating the random permutation data set, unbiased Fisher-Yates shuffle is applied so that every permutation is equally likely. And it is also efficient with the time complexity of *O(n)*. For area-based data, it shuffles every single person for each flow rather than flow numbers among different origins and destinations. And this shuffle could guarantee that both of inflow and outflow marginal for each area are unchanged.

Given n flows in the flow list, for i = n-1 to 1
generate a random integer r from 0 to i (inclusive)
exchange the destinations of flow(r) and flow(i)

Two things should be noted here. Firstly, a flow, which is selected before, needs to be exchanged with a randomly selected flow following the procedure. Secondly, a flow could be picked by itself when exchanging. This would grantee an unbiased random permutation.

3.4.2.2 FLOW TUBES

To construct flow tubes to scan the entire flow map, all the ordered pairs of geographical areas are iterated, and the tube size is varied according to user's settings. For example, in the case study of county-to-county migration, it uses all the directed flow tube with origin and destination regions covering one to five million populations. In other words, for each order pair of areas *(o, d)*, where *o* is flow origin and *d* is flow destination, it finds a region with one, two, three, four or five million populations for *o* and *d*, respectively. Therefore for each pair of counties, 25 flow tubes are created. If there are 1,000 geographical areas in the studied area, $1000 \times 999 \times 25 = 24975000$ potential flow tubes could be constructed to scan the flow map and each of them could be considered as a flow cluster candidate.

The algorithm to find a region covering a pre-defined population threshold is:

Step 1. Given an area A, find its first-order neighbors, put in a list L.
Step 2. Initialize P as the population size of a.
Step 3. Order the list L by the distance from a.
Step 4. For each area b in the ordered list L,
p = p + population size of b
if p > pop threshold, break for loop
Step 5. if p < pop threshold, extend searching list to the next-order neighbors and go back step 3.

For point-based data, thresholds for inflow and outflow could be used to control flow tube size. The reasons why I use regions covering at least one million populations inside for county-to-county migration data are: 1) unlike the cases which investigators use disease scan statistic to identify small disease clusters, researchers studying people migration are more interested in discovering significant patterns among relatively large area; 2) it is convenient to compare with other researches, which could help validate the

results; 3) it will dramatically reduce computational complexity. Users are free to define their flow tube sizes in terms of their research questions and data characteristics.

3.5 Data and Results

I evaluate the new method by investigating area-based flow data, a migration data set of Census 2000, and point-based flow data, a synthetic flow data which was used in (Guo & Zhu 2014). The migration data is aggregated at the county-level and considered as areal flow data. The synthetic point-to-point flow data is used to demonstrate and evaluate the flow scan statistic method for point-based flow data.

3.5.1 U.S. Internal County-to-county Migration

The Census 2000 migration data used in this study are among 3,075 counties in domestic U.S. Census 2000 covering a five year period of 1995 – 2000, which asked where the person lived five years ago (i.e. April 1, 1995). From 1995 to 2000, for the 3,075 counties analyzed in this study, there are 46,629,023 migrants crossing county boundaries, and 721,433 pairs of counties with non-zero migration flow. The values of net migration (the amount of immigrants minus emigrants) range from -568,788 to 212,235, and the net migration rate per 1,000 residents (net migration divided by amount of residents multiplying 1000) range from -442 to 300. Figure 3.4 presents the top 10,000 migration flows.

In addition to the whole dataset, I also explore the migration patterns for the older population (i.e. aged 65 and over). There are 3,150,152 older migrants moving among 151,566 pairs of counties. The values of net migration range from -46,468 to 26,157, and the net migration rate range from -1367 to 385.

For both data sets the proposed flow scan statistic is applied with a combination of one to five million flow tube sizes to scan the whole studied area. In other way, the sizes of flow tubes include 25 options, from one of 1-5 million areas to another one of 1-5 million areas. In the process of *reporting* significant flow clusters, only the clusters which do not spatially overlap with the significant clusters of higher statistic values are reported. "Spatially overlapping" for flow tubes in this study is defined as spatial overlap occurring on both origins and destinations at the same time. What's more, the migration clusters with distance less than 300 kilometers are excluded because the migration with really short distance is much more trivial than long-distance migrations and could mess the map. Table 3.1 summarizes the count of flow clusters under different criteria.

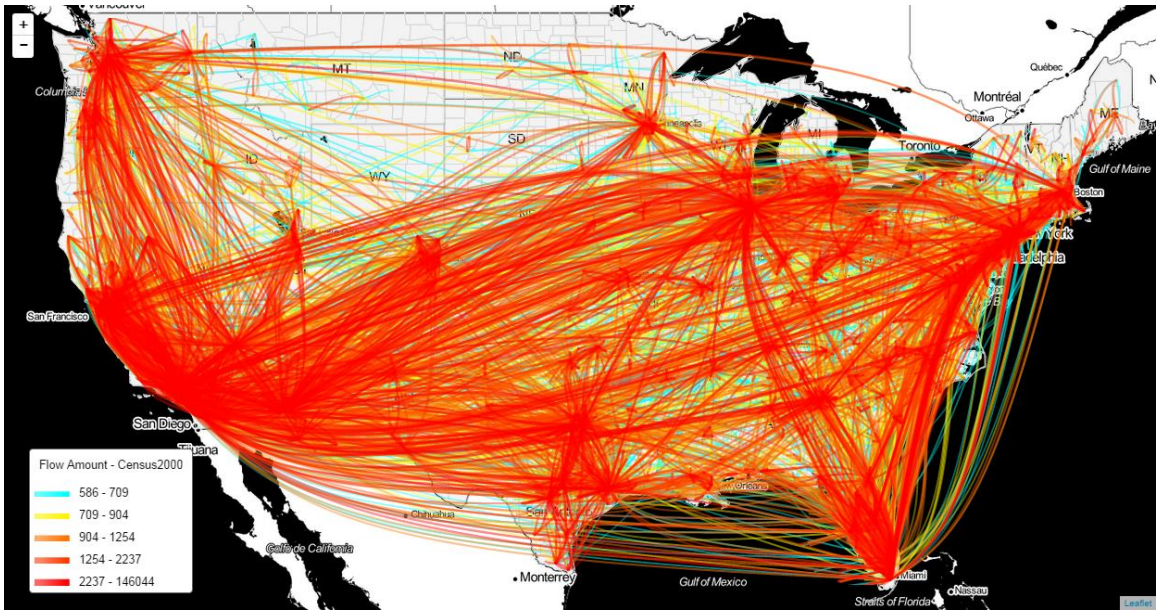


Figure 3.4: Top 10,000 migration flows out of 721,433 among 3075 counties for Census 2000. The breaks are classified by flow amounts.

Figure 3.5 presents 307 significant migration flow clusters with distance larger than 300 kilometers at the significant level 0.001 for the entire population. Although all

the presented flow clusters are significant, they are classified based on the value of *LGLR* using quantile classification and assigned with different colors. Generally obvious migration patterns are hot flows from Greater New York area to Florida and strong interactions within Texas and California. Figure 3.6 shows 120 flow clusters excluding double-sided flows in Figure 3.5, which means that it excludes the pairs of flows one of whose origin and destination overlap another's destination and origin, respectively. The map clearly shows the migration flow patterns, especially within Texas and California. It also suggests that there is strong outflow trend from Florida to Atlanta area although Florida is still a hot destination for former Northeastern residents.

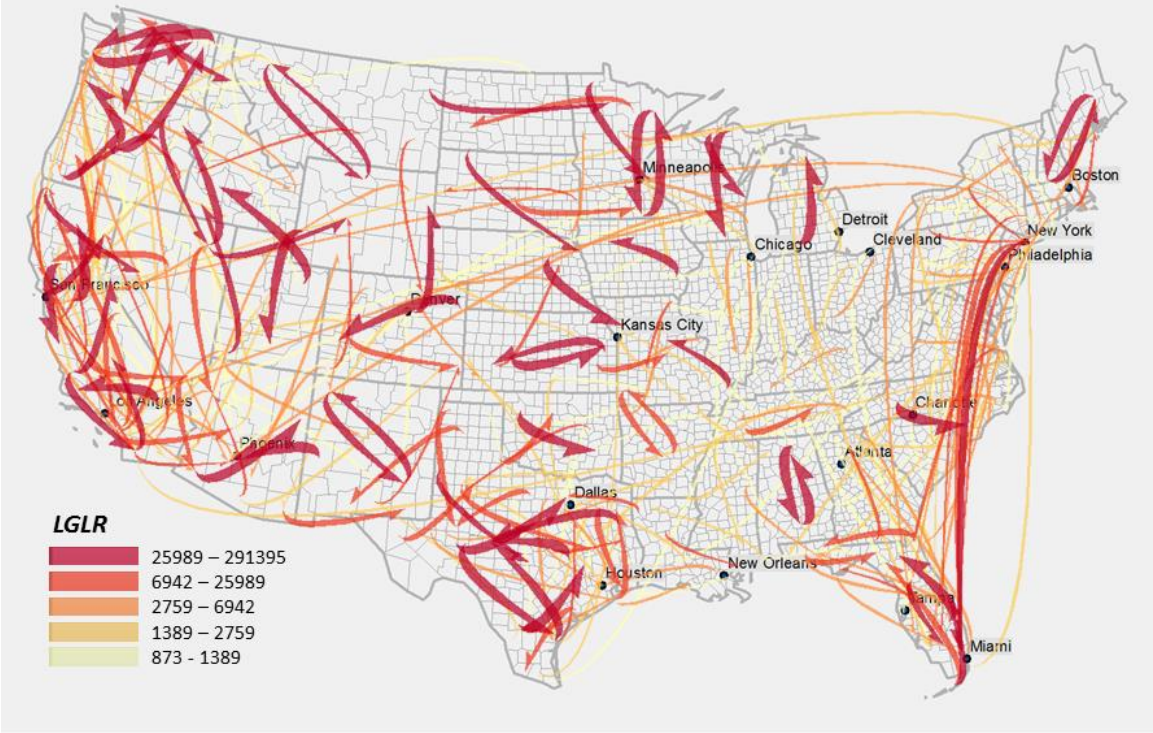


Figure 3.5: Significant migration flows for entire population (*p-value* < 0.001).

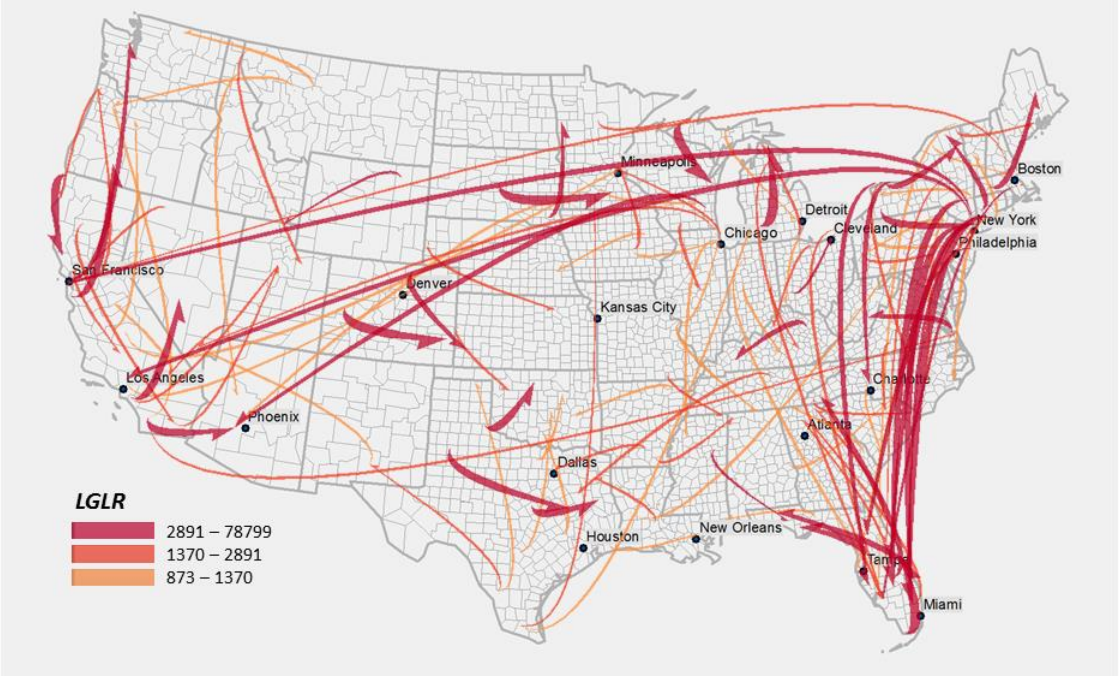


Figure 3.6: Significant migration flows without double-sided directions for entire population (p-value < 0.001)

Figure 3.7 shows the analyzed migration patterns for the older population, and Figure 3.8 shows the no-double-sided flows. Generally, a strong migration trend from north to south can be observed. Significantly strong migration flows are from Northeast Region to Florida area. Arizona is the hot destination for the movers from West and Midwest U.S. Texas is attractive to the movers from Midwest U.S. as well. In general, the results has discovered the significant hot flow patterns as expected and those patterns are also evidenced in (Guo & Zhu 2014).

Table 3.1: Counts of flow clusters under different criteria.

	Criteria	Entire pop	Older pop
(1)	All non-overlapped flow clusters	2472	3081
(2)	(1) and *LGLR* > cut-off value (significant)	785	751
(3)	(2) and Length of flow clusters > 300km	307	335
(4)	(3) and No double-sided direction	120	157

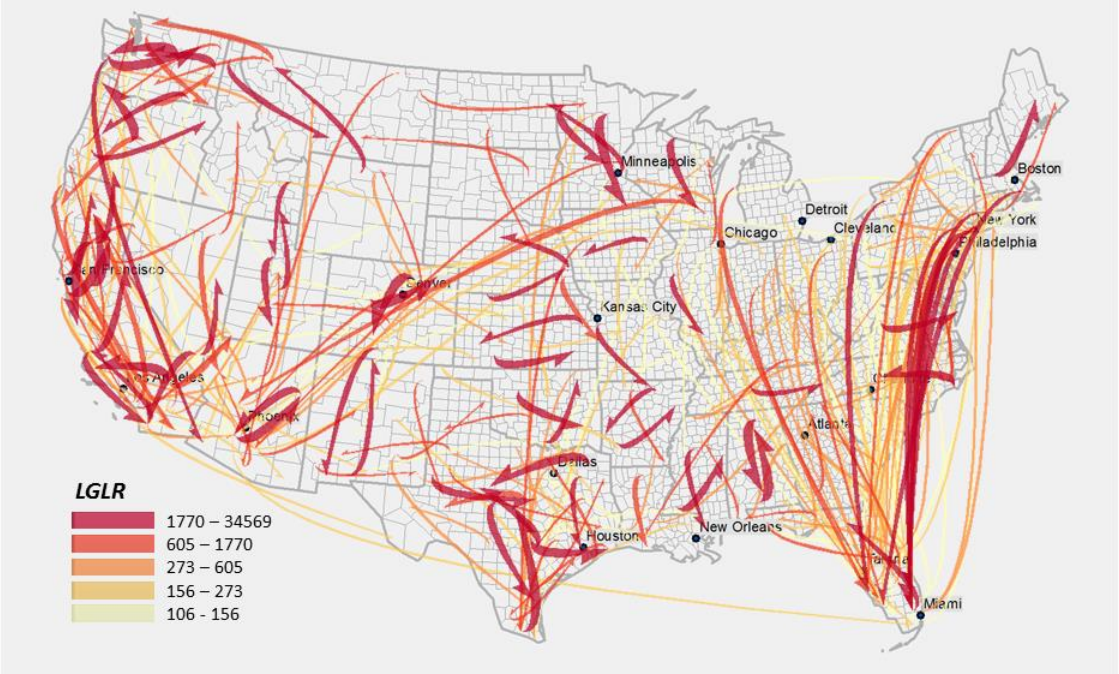


Figure 3.7: Significant migration flows for age above 65 in Census 2000 (*p-value* < 0.001).

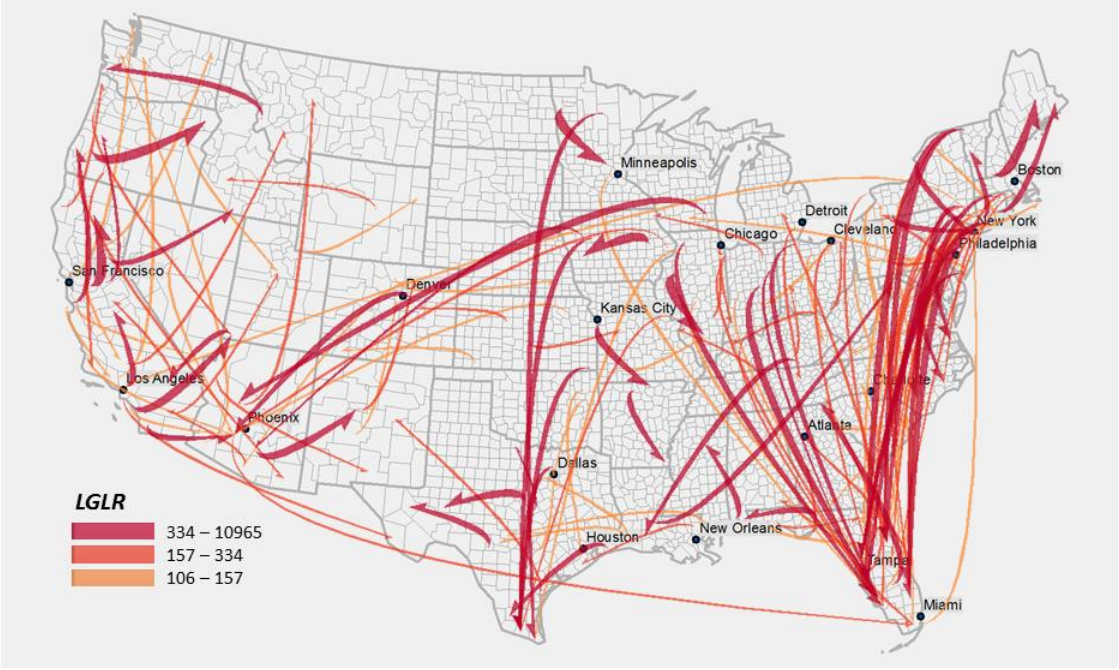


Figure 3.8: Significant migration flows without double-sided directions for age above 65 (*p*-value < 0.001).

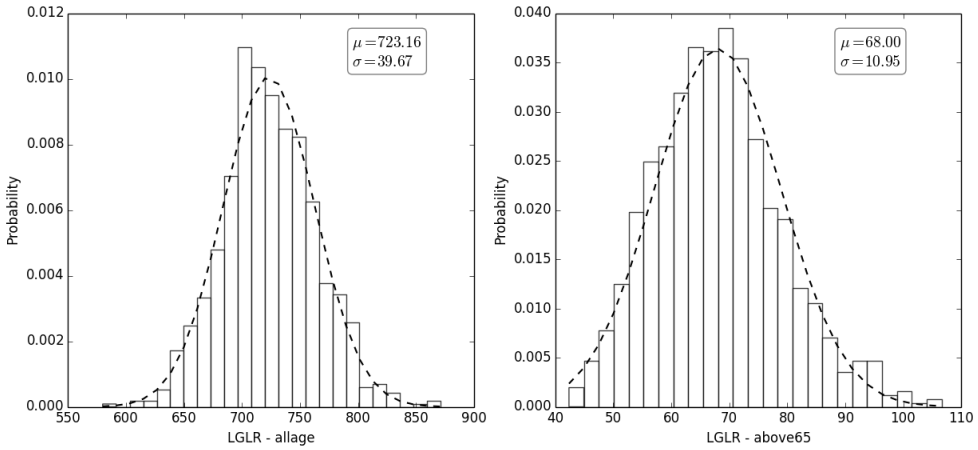


Figure 3.9: Histograms of *LGLR* values under null hypothesis for entire migrants (left) and older migrants (right). The dotted curves are normal distribution with mean and standard deviation of *LGLR*.

To examine multiple-testing problem of flow scan statistic, the adjusted cut-off points are investigated using Bonferroni adjustment. By examining the histograms and their fitness of normal distribution of random *LGLR* values in Figure 3.9, it is assumed that *LGLR* values follow a normal distribution. Therefore, the adjusted cut-off points are 902.07 and 920.57 at significance level 0.01 and 0.001 for entire migration data, and 117.40 and 122.50 at significance level 0.01 and 0.001 for older migration data, respectively. Comparing with the cut-off points shown in Figure 3.5 and Figure 3.7, adjusted cut-off points would not make big difference.

For the old population data, average computational time for one permutation including random data generation (1.016 ± 0.054 seconds) and flow scanning (2.138 ± 0.142 seconds) is 3.153 ± 0.174 seconds. In total, about 53 minutes are taken to scan the flows and to perform 999 Monte Carlo simulations. An Intel Core i3 3.06GHz processor with 12 GB RAM and Windows 7 was used. Performance would be dramatically improved by taking advantage of parallel computing.

3.5.2 POINT-BASED FLOW DATA - SYNTHETIC DATA SET

To evaluate the performance on point-based flow data, I apply flow scan statistic on synthetic flow data set which was generated in (Guo & Zhu 2014). The process of synthetic data generation is as follows:

1. *Divide the studied area into 100*100 pixels, generate base population for each pixel with two designed urban areas (upper right and lower left);*
2. *Based on the base pop, generate (#clustersize*7*10 =) 3500 points as origins and the other 3500 points as destinations.*
3. *For each designed cluster, selected nearest #clustersize/density points for origins and destinations, respectively. From those points, randomly select #clustersize origins and destinations to construct flows, and those are assigned as flow clusters.*
4. *For the rest points, randomly select two to make a flow until all the points are assigned.*

Table 3.2: Seven designed point-based flow clusters.

	Origin	Destination	Cluster Size	Density	Base size	Expected
Blue	(200,200)	(800,800)	50	0.5	100	2.86
Green	(900,400)	(200,100)	50	0.5	100	2.86
Pink	(900,300)	(200,100)	50	0.25	200	11.43
Yellow	(200,200)	(900,800)	50	0.15	333	31.68
Red	(300,400)	(300,400)	50	0.1	500	71.43
Magenta	(800,200)	(400,900)	50	0.1	500	71.43
Cyan	(300,800)	(800,700)	50	0.1	500	71.43

The seven designed flow clusters are described in

Table 3.2. All seven flow clusters contain 50 movers, but cluster densities vary. Blue and green clusters have the highest density of 0.5, which means that at least 50% of the nearest 100 points of origin is moving to the nearest 100 points of destination. Pink cluster has second highest density, and yellow cluster has moderate density. Red, magenta and cyan clusters have the lowest density, and the expected flow amount is higher than the designed value, in which case it is a challenge to detect those as significant clusters. The red cluster is designed as noise since the origin and destination

are the same. The clusters are also designed with geographic meaning in mind. Blue and yellow clusters are intended to simulate flows between urban areas, while green and pink are intended to simulate flows between urban and rural areas. Magenta and cyan are between rural areas.

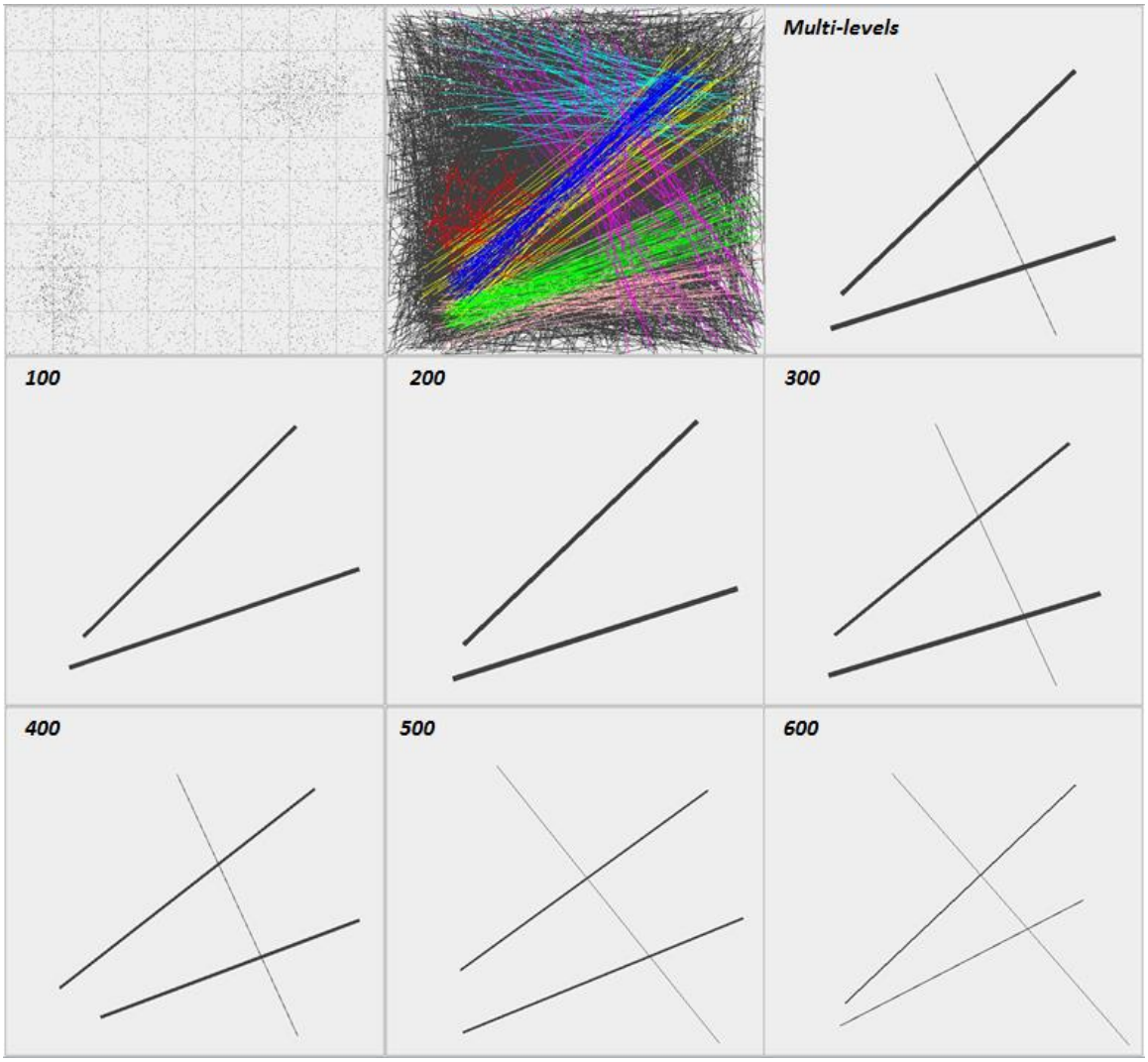


Figure 3.10: Synthetic point-based flow data and flow scan results. From top left to bottom right: 1) 7000 randomly points given two urban areas; 2) 3500 flows with seven designed clusters; 3) flow clusters with *p-value* < 0.05 using multi-level scanning strategy (i.e. nearest 100, 200, 300, 400, 500 and 600 points); 3) to 9) are flow clusters with *p-value* < 0.05 using nearest 100, 200, 300, 400, 500 and 600 points, respectively. In figure 3) to 9), the line width represents the *log*(*GLR*) value.

To scan point-based data, the two bases of flow tube (origin and destination) are designed as certain regular shape (e.g. circle and etc.) covering certain amount of flows. The parameter controlling tube size varies in origin and destination, which contributes to flow tubes with different base shapes and sizes. The amounts of inflow and outflow are used to control the tube size instead of population. Flow scan statistic is applied with multiple flow tube sizes from nearest 100 to 600 points (36 size combinations in total). To test the sensitivity of flow tube size, it is also with six constant tube sizes (i.e. 100, 200, 300, 400, 500, and 600). The data and results are presented in Figure 3.10. The approach has detected green and pink, blue and yellow clusters for all choices of flow tube sizes, although it does not discriminate green from pink and blue from yellow. For magenta cluster, multi-level flow tube sizes could recognize it although the p-value is 0.017 (as shown in Table 3.3). For the constant tube size results, when the size increases to 300, magenta cluster was detected.

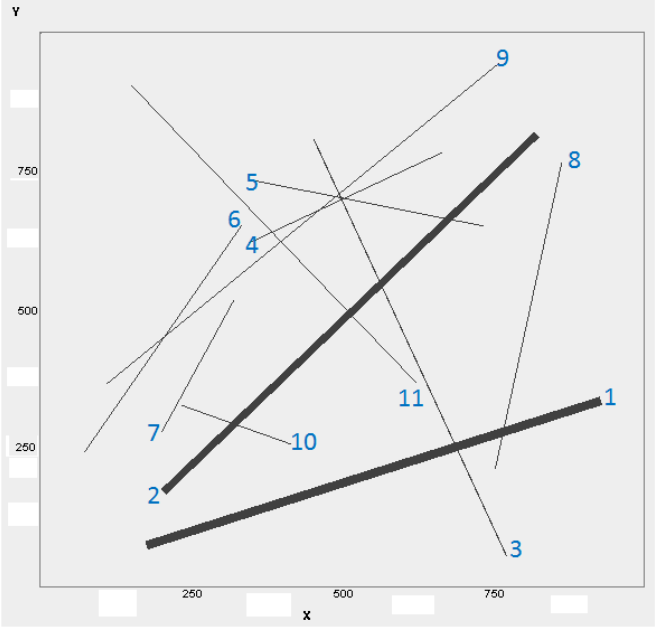


Figure 3.11: Top flow clusters using multi-level flow scan statistics. Starting areas of flow clusters are marked by cluster numbers. Details of flow clusters refer to Table 3.3.

The multi-level flow clusters with positive *log(GLR)* values are presented in Figure 3.11 and Table 3.3. The sizes of detected clusters matched the designed sizes. Green and pink clusters are detected as the most significant cluster with 200 origin and destination size. The second significant cluster is blue and yellow clusters, whose origin and destination size are 200 as well. The third detected cluster is magenta cluster with 400 as both origin and destination size. Although it was not detected as significant, cyan cluster was ranked as the fifth highest *log(GLR)* value in the result.

Table 3.3: List of all reported flow clusters. See Figure 3.11 for related maps.

	Origin	**Destination**	**Origin size**	**Destination size**	**log(GLR)**	***p*-value**
1	(922, 332)	(183, 77)	200	200	78.11	0.001
2	(208, 176)	(819, 812)	200	200	65.16	0.001
3	(773, 56)	(453, 808)	400	400	10.65	0.017
4	(345, 621)	(665, 784)	300	100	4.71	0.999
5	(355, 733)	(732, 652)	400	100	4.53	0.999
6	(332, 652)	(72, 245)	500	400	2.87	0.999
7	(201, 280)	(319, 517)	300	100	1.97	0.999
8	(863, 765)	(755, 213)	600	400	1.87	0.999
9	(759, 945)	(110, 366)	500	400	1.59	0.999
10	(414, 259)	(235, 328)	100	100	1.31	0.999
11	(621, 370)	(151, 904)	600	100	1.22	0.999

3.6 Discussion and Future work

In this Chapter I presented a new flow scan statistic method for detecting significant flows clusters, at different scales, from spatial interaction data. The test statistic is based on a *Generalized Likelihood Ratio (GLR)*. A large number of flow tubes are constructed to scan the entire study area and flow data. Users could define their own construction rules to control the tube size. Monte Carlo permutation is employed to obtain the distribution of the test statistics under the null hypothesis. Based on the null

distribution obtained from the permutation, a *p*-value can be assigned to each flow cluster.

The internal county-to-county migration data in continental U.S. from 1995 to 2000 are used to evaluate the methodology. Migration patterns for all migrants and for senior population are analyzed. The results show that the approach can efficiently uncover migration flow trends from massive flow data, for a large geographical area, and at different scales (controlled by the neighborhood size of flow tubes). In addition to area-based flow data (such as county-to-county migration), the flow scan statistic method can also work with point-based flow data, which was demonstrated with a synthetic data. The results show that the proposed approach is effective in detecting significant flow clusters with robust performance as demonstrated by different experiments and parameter settings.

In this study a flow tube is defined with regular-shaped neighborhoods (i.e. circles) for its origin and destination. Future extensions can explore the use of other neighborhood definitions to achieve more accurate results, such as integrating certain regionalization or clustering methods to capture inherent characteristics of inflows and outflows. Moreover, since spatial flow patterns often change over time, further work is needed to extend the method to detect spatial-temporal flow clusters and trends.

CHAPTER 4： AN EXPLORATORY APPROACH TO SPATIAL INTERACTION MODELING AND RESIDUAL ANALYSIS

4.1 ABSTRACT

Spatial interaction models have been studied extensively for analysing a variety of geographic mobility data including migration. The design of spatial interaction models usually follows a *confirmatory* process that consists of the design of a model and the interpretation of its configured parameters with observational data. This research presents an *exploratory* framework to analyse and map the residuals of fitted spatial interaction model. The residuals can detect important patterns that the model cannot explain. By coupling model and the spatial analysis of its residuals the reader can better understand not only the global trend captured by the model but also the unknown patterns behind the model such as spatial association and clustering (local migration patterns).

I extend the local Moran's *I* statistic to quantify the spatial clustering of flow prediction residuals (which are also flows) in a spatial interaction model. A case study is developed using the Census 2000 county-to-county migration data in the U.S., particularly focusing on the migration flows among 358 Metropolitan Statistical Areas (MSAs). The migration flow between two MSAs is further stratified into seven age groups. A series of analysis are carried out to configure a spatial interaction model and analyze its outcomes to obtain a variety of patterns and insights for each age group.

The overall contribution of this research is threefold. First, a new variant of spatial interaction model is developed. Second, a new flow-based spatial autocorrelation statistic is designed to analyze and map model residuals. Third, with the developed approaches (modelling and residual mapping), new insights and patterns in the migration data are discovered, which existing methods cannot find.

4.2 INTRODUCTION

There are extensive studies on spatial interaction models that aim to analyze, understand and predict migration and other geographic mobility (Fotheringham et al. 2000, Roy & Thill 2004, Chun et al. 2012). Existing research on spatial interaction modeling usually follows a *theory-driven* and *confirmatory* framework, which constructs a model based on theoretical knowledge and fits the model with observational data to confirm its power and understand its parameters. While existing research mostly focus on interpreting model parameters and examining global patterns, there is much less attention given to the analysis of model residuals and the discovery of unknown local patterns that are masked by the global trend.

The research presented in this paper captures global trends with a modeling approach, explores the residuals of the fitted spatial interaction model, and discovers unknown local patterns that the global model cannot explain. I analyze the Census 2000 migration data set, which is aggregated by 358 Metropolitan Statistical Areas and seven age groups. The aim is to (1) configure a model to capture the global trend of U.S. internal migration in relation to well-known factors such as population, distance, and competing destinations; and (2) map and analyze the model residuals to discover migration patterns not explained by the model (such as significant larger migration flows

than predicted) and regions that exhibit different local migration patterns from the global trend.

Specifically, there are three contributions in my research presented in this paper. *First*, among various gravity models for migration analysis, through exploratory analysis I find a piecewise version of the power-law gravity model that is more suitable for the U.S. migration analysis. The configured models find that the negative impacts of geographic distance and competing destinations are much stronger on short-range migration than impacts on long-distance migration (except for age group above 60), with significant distance breakpoints. *Second*, the local Moran's *I* statistic is extended to measure the spatial clustering of migration prediction residuals (note: each residual is a net residual flow between two locations). Through residual mapping and statistical analyses, it discovers a variety of interesting migration patterns deviated from the global trend captured by the model. With the model and the residual analysis, one can obtain a more comprehensive understanding of hidden migration patterns that vary from place to place. *Third*, the exploratory analysis of residuals can provide insights for model improvement and understanding complex local patterns. For example, it is found that the migration patterns in age groups 05-14 and 30-44, age groups 15-19 and 20-24, and age groups 45-59 and above 60 share similar patterns in net- residual maps, suggesting new categorization of age groups for migration analysis.

4.3 BACKGROUND

4.3.1 GRAVITY MODEL

Spatial interaction modeling has long been studied and a variety of spatial interaction models have been developed. One of the earliest and commonly used model types is the *gravity model*, which assumes that the flow from one location to another is proportional to their mass — a measure of "importance" for each location — and inversely proportional to their distance or any other measure of "friction". Following are two formulations of gravity models, which differ in their assumed relationship between the flow (F_{ij}) and geographic distance (D_{ij}):

$$F_{ij} = \alpha {P_i}^{\beta} {P_j}^{\gamma} e^{\theta D_{ij}} \quad \text{or}$$

$$\ln F_{ij} = \ln \alpha + \beta \ln P_i + \gamma \ln P_j + \theta D_{ij} \quad (4.1)$$

$$F_{ij} = \alpha {P_i}^{\beta} {P_j}^{\gamma} {D_{ij}}^{\theta} \quad \text{or}$$

$$\ln F_{ij} = \ln \alpha + \beta \ln P_i + \gamma \ln P_j + \theta \ln D_{ij} \quad (4.2)$$

where P_i and P_j are the population (or other attributes) of the origin i and destination j. In Equation 4.1 flow increases exponentially with distance; Equation 4.2 represents a power-law relationship. Both models are commonly used in geographic mobility analysis. For example, the former is used in a recent study on commodity flows (Chun et al. 2012) and the latter is applied in analyzing state-to-state migration in the U.S. (Chun 2008). Socio-economic network studies often prefer the power-law model. For example, a power-law model for mobile communication networks showed that the exponent for distance is close to 2 (Lambiotte et al. 2008). Kaluza et al (2010) studied ship movements by integrating a truncated power law model into the gravity model. Viboud and others

(2006) estimated the parameters for a piecewise gravity model fitted to U.S. workflow data by county.

Gravity models have been used in many empirical network applications (Barthélemy 2011). Jung, Wang and Stanley (2008) studied the traffic flows of Korean highway system as a proxy of human mobility. For 30 cities with population over 200,000, they found that the traffic flows formed the originate gravity model with the formula as

$$F_{ij} \sim P_i P_j {d_{ij}}^{-2}. \tag{4.3}$$

Kaluza et al. (2010) applied a gravity model in the global cargo ship network. The specific model being used is a truncated power law function, defined as

$$F_{ij} = A_i B_j O_i I_j {d_{ij}}^{-\rho} e^{-d_{ij}/\gamma}, \tag{4.4}$$

where O_i is the total number of departure records in port i, I_j is the total number of arrival records in port j, A_i and $B_j$are constraint coefficients ensuring $\sum_i F_{ij} = O_i$ and $\sum_j F_{ij} = I_j$.The strongest correlation was obtained for ρ=0.59 and γ=4900 km.

Balcan et al. (Balcan et al. 2009) analyzed commuting data from 29 countries between subpopulation areas defined by a Voronoi decomposition. They found that the best fit was obtained by using exponential laws and including a characteristic length governing the decay of commuting flows in the deterrence function, shown in:

$$F_{ij} = k P_i^{\alpha} P_j^{\beta} e^{-{d_{ij}}/{r}}, \tag{4.5}$$

where P_i and P_j are the population of subpopulation area i and j, respectively. The estimated values for α, β, and γ are 0.46, 0.64, and 82 when $d_{ij} \leq 300$ km, and 0.35, 0.37, and NA (Not Available) when $d_{ij} > 300$ km. At a smaller scale, Viboud et al. (2006)

estimated the parameters for the piecewise gravity model fitted to continental U.S. work flow data between 3109 counties and found a breakpoint of 119 km. The form of gravity model they used was:

$$F_{ij} = kP_i^{\alpha} P_j^{\beta} {d_{ij}}^{-\rho}. \quad (4.6)$$

The estimated values for α, β, and ρ are 0.30, 0.64, and 3.05 when $d_{ij} < 119$ km, and 0.24, 0.14, and 0.29 when $d_{ij} \geq 119$ km. Different results might have originated in the different scales used in these two studies. Viboud et al. (2006) used county boundaries, while Balcan et al. (2009) used more statistically homogeneous subpopulation area defined by a Voronio decomposition.

Table 4.4: A survey of empirical studies with gravity models (Source: Barthélemy 2011)

	#node	Gravity model form	Results
Korean highway (Jung et al. 2008)	30	$P_i P_j {d_{ij}}^{-\rho}$	ρ=2
Global cargo ship (Kaluza et al. 2010)	951	$A_i B_j O_i I_j {d_{ij}}^{-\rho} e^{-d_{ij}/\gamma}$	ρ=0.59, γ=4900 km
Worldwide commuting (Balcan et al. 2009)	NA	$kP_i^{\alpha} P_j^{\beta} e^{-d_{ij}/r}$	(α, β, γ) = (0.30, 0.64, 3.05) with R^2 = 0.80 when $d_{ij} \leq 300$ km; (α, β, γ) = (*0.35, 0.37, NA*) with R^2 = 0.54 when $d_{ij} > 300$ km
Continental U.S. commuting by county (Viboud et al. 2006)	3109	$kP_i^{\alpha} P_j^{\beta} {d_{ij}}^{-\rho}$	(α, β, ρ) = (0.46, 0.64, 82) when $d_{ij} \leq 119$ km; (α, β, ρ) = (0.24, 0.14, 0.29) when $d_{ij} > 119$ km

Many researches attempted to improve the performance of the gravity model for estimating flows. Karemera, Oguledo and Davis (2000) adopted an empirical gravity model which included a set of location variables representing adjacency, population density, and dummy variables of six regions. Specification tests were conducted to determine the significance of each additional variable. Goh et al. (2012) modified the

gravity model by integrating the Hill function, which is widely used for chemical reactions, to deal with the cut-off behavior of the power-law distribution, and applied it to the passenger flows in the Metropolitan Seoul Subway system. Vitali and Battiston (2011) discussed the usage of geographical and network/graphical distance in the gravity model. Fischer and Griffith (2008) added a technological distance in the gravity model. A summary of the recent and representative studies on the gravity model is shown in Table 4.4. The choice of deterrence functions, mass variables, and the estimated model parameters are different from case to case. Different scales are used in the studies surveyed.

There are numerous variants and extensions for gravity models, such as the competing-destination model (Fotheringham 1983, Fotheringham 1986), intervening-opportunity model (Stouffer 1960), entropy-based models (Wilson 1967, Roy & Thill 2004), neural networks (Fischer & Gopal 1994, Openshaw 1998, Fischer et al. 2003), models based on complex network theories (Andersson et al. 2006), and models that take into account the spatial autocorrelation among flows (Chun et al. 2012, Chun 2008). Extensions of the gravity model usually include dummy variables to represent different regions/counties (Karemera et al. 2000, Cohen et al. 2008) or a competing destination (*CD*) variable to measure the accessibility of a destination *j* to all other destinations (Fotheringham 1983, Fotheringham 1986, Guldmann 1999). In this research, I integrate a *CD* variable in the models shown in Equations 4.1 and 4.2. Specifically, a *CD* value is calculated for each location *i* with the following formula (Equation 4.7), and the gravity model is reformulated as Equation 4.8.

$$CD_i = \sum_j P_j / D_{ij}\,, \quad i \neq j \tag{4.7}$$

$$F_{ij} = \alpha P_i^{\beta} P_j^{\gamma} {D_{ij}}^{\theta} CD_j^{\rho} \tag{4.8}$$

4.3.2 ALTERNATIVE MODELS

In addition to gravity models, there are also other models that have been developed to estimate interactions. Recently, Simini et al. (2012) proposed a radiation model for generating commuting networks. It was based on two simple assumptions: people do not like move, and they move to the nearest opportunity that could improve their life. The radiation model could be formulated as:

$$F_{ij} = \left(P_i \frac{F}{P}\right) \frac{P_i P_j}{(P_i + P_{ij})(P_i + P_j + P_{ij})}, \tag{4.9}$$

where F is the total number of commuters and P is the total population in the studying area, P_{ij} is the total population in the circle of radius d_{ij} centered at i (excluding P_i).

Table 4.5: Parameter estimates of the piecewise gravity model fitted to continental U.S. commuting data by 3109 counties. Models are fitted separated for distance above and below 129 km, which is chosen as the cut-off point that minimized the residual sum of square of a piecewise gravity model. The adjusted R^2 for the whole model is 0.6346.

	d_{ab} < 129 km		d_{ab} > 129 km	
	Coeff.	Std. error	Coeff.	Std. error
Intercept	6.41***	0.068	-1.12***	0.026
$ln(P_i)$	0.34***	0.004	0.28***	0.002
$ln(P_j)$	0.66***	0.004	0.15***	0.002
$ln(d_{ij})$	3.12***	0.012	0.31***	0.003
Adjusted R^2	0.6375		0.2557	

***: p-value < 0.001

Although this study compared with the results of gravity model provided in (Viboud et al. 2006), it seemed that they made a mistake by ignoring the constant

parameter *k* which was not provided in Viboud's published paper, or misunderstood the piecewise regression. I use the same data and gravity model form, and obtain similar results as in (Viboud et al. 2006) shown in Table 4.5. I calculate the estimated commuting flow originating from New York City and show the results in Figure 4.1. It is clear that the gravity model gives more realistic approximation to the observed commuting patterns.

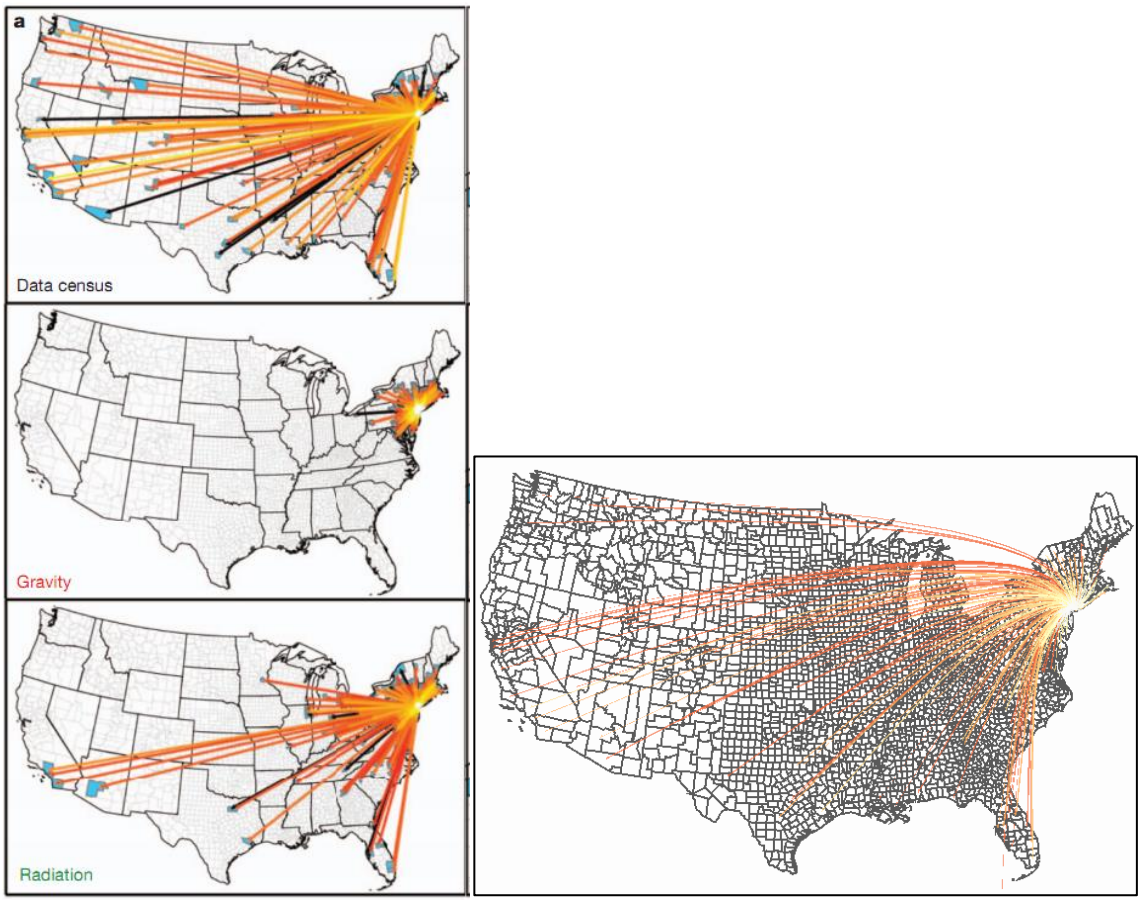


Figure 4.1: Results in (Simini et al. 2012) and results of my gravity model. Flows with less than ten individuals are not shown. Upper-left shows the national commuting flows originating from New York County. Middle-left is the predicted results using a gravity model in (Simini et al. 2012), which I believe is wrong. Bottom-left is the result of a radiation model in (Simini et al. 2012). The map on the right column displays my results with a gravity model.

Lenormand et al. (2012) proposed an individual based procedure according to a probability P_{ij}, which is inspired by gravity models, to build the simulated commuting networks. Only a single parameter β in the probability function, which rules the compromise between the influence of the distance and job opportunities, is estimated by minimizing the Kolmogorov-Smirnov (KS) distance between the observed and simulated distribution of commuting distances. The simulated networks are generated by considering each single commuter's choice for its place of work.

$$P_{ij} = \frac{F_j \exp(-\beta D_{ij})}{\sum_{k=1}^{N} F_k \exp(-\beta D_{ik})}, \tag{4.10}$$

where P_{ij} is the probability for a commuter from unit *i* to unit *j*, and F_j is total number of commuters entering in unit *j*. By testing the model on 80 cases with geographic units of different sizes, the parameter β was found very relevant to the unit area of studying case. Many studies have shown that piecewise regression yields improved result, since distances (long/short) have significantly different impact on flow interaction. However, almost all the case studies in this paper, except for one case for county level commuters in the U.S., are based on small study area. It has become a critical challenge to extract meaningful patterns from increasingly large and complex flow data.

4.3.3 Evaluation Measures

To evaluate and compare different spatial interaction model outcomes, several measures have been suggested in the literature (Knudsen & Fotheringham 1986, Hu & Pooler 2002, Thorsen & Gitlesen 1998). In this research, I adopt four commonly used measures, including percentage misallocated (*PM*), standardized root mean square error (*SRMSE*), the *PSI* statistics, and the *log-likelihood* value in Poisson regression for

evaluation. The *PM* measure represents the percentage of flows that are misallocated in the flow matrix (Hu and Pooler 2002), defined as:

$$PM = \frac{50}{F}\sum_i\sum_j |f_{ij} - \hat{f}_{ij}| \tag{4.11}$$

where f_{ij} is the observed flow, $\hat{f}_{ij}$ is the predicted flow volume, and *F* is the total flow volume in the data. *SRMSE* is a general distance between f_{ij} and $\hat{f}_{ij}$, defined in (Pitfield 1978):

$$SRMSE = \frac{[\sum_i \sum_j (f_{ij} - \hat{f}_{ij})^2/N]^{0.5}}{\sum_i \sum_j f_{ij}/N} \tag{4.12}$$

where *N* is the total number (note: not volume) of flows (e.g., there are at most 100*99 = 9900 flows with 100 regions). The *PSI* measure is an information-based statistics introduced in (Ayeni & Referee J.B.H. Ramsey 1983):

$$\Psi = \sum_i\sum_j p_{ij}|\ln(p_{ij}/s_{ij})| + \sum_i\sum_j q_{ij}|\ln(q_{ij}/s_{ij})| \tag{4.13}$$

where $p_{ij} = f_{ij}/\sum_i\sum_j f_{ij}$, $q_{ij} = \hat{f}_{ij}/\sum_i\sum_j \hat{f}_{ij}$, and $s_{ij} = (p_{ij} + q_{ij})/2$. All three measures, i.e., PM, SRMSE and PSI, have smaller values for better prediction models. The log-likelihood measure that is maximized in the Poisson regression procedure is also used in this research as an indicator of model performance. To quantify the prediction residual for each individual observation (e.g., flow), I use the *Standardized Deviance Residual*, which is widely used in Poisson regression (McCullagh & Nelder 1989):

$$r_i = sign(y_i - \hat{y}_i)\sqrt{2[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - y_i + \hat{y}_i]}/\sqrt{1 - h_{ii}} \tag{4.14}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix *H*, as used in Poisson regression (McCullagh & Nelder 1989). Conceptually *Standardized Deviance Residual* is more

comparable than regular residual because it takes into account different variances at different observations and additional variation deriving from parameter estimations.

4.3.4 Network Autocorrelation

Spatial autocorrelation quantitatively assesses the degree to which the value of one random variable at a specific location is dependent on the values at its neighboring locations. Studies on spatial autocorrelation have been conducted for several decades (Moran 1948, Anselin 1995, Getis 2008, Geary 1954, Getis & Ord 1992, Ord & Getis 1995). Moran's *I* and Geary's *C* are the most commonly used statistical indices for global autocorrelation analysis, and extended versions of these two statistics were also proposed to detect local autocorrelation(Anselin 1995).

The existence of network autocorrelation has also been recognized and analyzed in several studies (Black 1992, Black & Thomas 1998). Black (1992) defined network autocorrelation as the influence of the variables associated with a link on its interconnected links. Black (1992) developed an extension of spatial autocorrelation analysis statistic Global Moran's *I* for SI data, and Black and Thomas (1998) demonstrated the index by applying it to 1991 motor vehicle accident rates for a portion of the motorway network of Belgium. However, the global statistic can only tell us whether the network autocorrelation exists or not, but not the local information, e.g., hot flows and network outliers (analogous to the hot spots and spatial outliers in spatial autocorrelation). To eliminate this drawback, Berglund and Karlström (1999) presented a local statistic of network autocorrelation by generalizing the statistic of Getis-Ord G_{ij}. In analogy with the spatial version of the local G_i statistic, a statistically significant high value means that a flow with high value is surrounded by flows with similar high values

(hot flow), and a statistically significant low value means that a flow with low value is surrounded by flows with similar low values (cold flow). Berglund and Karlström (1999) used two different kinds of weight matrices defining the neighborhood of flows, and (Chun 2008b, Fischer & Griffith 2008, Chun et al. 2012b) added two more. The matrices are as following (illustrated in Figure 4.2 a, b, c, and d, respectively):

$$W_{ij,kl} = \begin{cases} 1 & if\ i = k\ and\ w_{jl} = 1 \\ 0 & otherwise \end{cases} \tag{4.15}$$

$$W_{ij,kl} = \begin{cases} 1 & if\ j = l\ and\ w_{ik} = 1 \\ 0 & otherwise \end{cases} \tag{4.16}$$

$$W_{ij,kl} = \begin{cases} 1 & if\ i = k\ and\ w_{jl} = 1, or\ if\ j = l\ and\ w_{ik} = 1 \\ 0 & otherwise \end{cases} \tag{4.17}$$

$$W_{ij,kl} = \begin{cases} 1 & if\ w_{ik} = 1\ and\ w_{jl} = 1 \\ 0 & otherwise \end{cases} \tag{4.18}$$

where w_{ij} denotes the spatial contiguous weight matrix which is commonly used in spatial autocorrelation analysis.

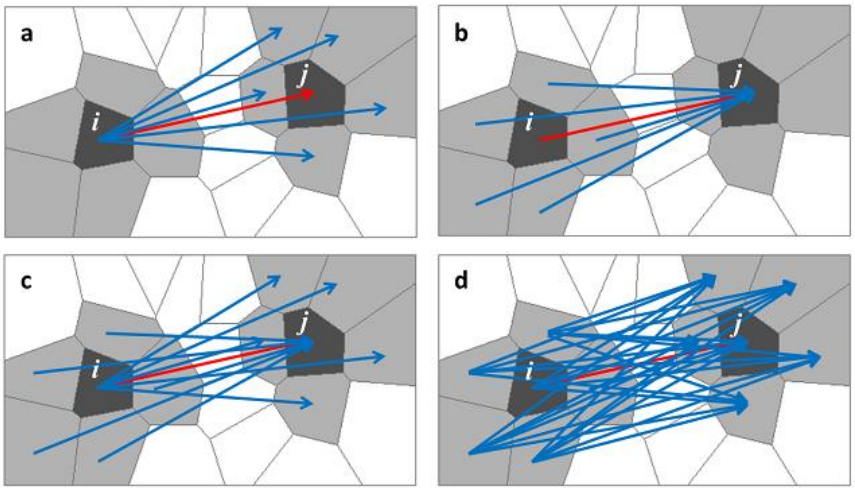


Figure 4.2: Illustration of weight matrices. Blue flows are the defined neighbors of the red flow according to corresponding weight matrices.

As mentioned before, local spatial statistic G_i is useful to detect the hot spots and cold spots. In recent research (Chun et al. 2012, Fischer & Griffith 2008, Chun 2008) network autocorrelation has been embedded into the data modeling by introducing a network error or appropriate synthetic surrogate variables (i.e. spatial filters).

4.3.5 FLOW MAP

Exploratory and visual approaches such as flow map have also been used to present spatial interaction data and patterns (Phan et al. 2005, Tobler 1976, Tobler 1987, Rae 2009, Guo 2009). Facing the challenges of complex and large spatial interaction data, visual approaches often rely on derived measures, spatial aggregation and user selections in order to focus on a certain aspect of flow patterns. The presented research makes contributions in two aspects of exploratory spatial interaction analysis. First, existing approaches often map the raw flows among units (e.g., Phan et al. 2005, Holten 2006, Holten & Wijk 2009), which are often dramatically different in size (in terms of population or area) and consequently the flows among them are not directly comparable. My approach models the flows by taking into account several well-known factors, such as population and distance, and then focuses on the analysis of residuals to discover unknown patterns and factors. Second, existing exploratory approaches for spatial interaction analysis often lack rigorous statistical testing and thus are unable to distinguish significant patterns from random data variations. I extend traditional local spatial statistic to test the significance of the spatial clustering and model residuals.

4.4 MIGRATION DATA

In this research I analyze the Census 2000 migration data for a five-year period of 1995—2000. Census 2000 asked where the person lived five years ago (i.e., April 1, 1995) and thus the data includes movers who moved within five years. The original migration data is at the county level. In this study, I focus on the migration among 358 metropolitan statistical areas (MSAs) within the continental U.S., including 48 states and Washington D.C. (Figure 4.3). There are 22,966,934 migrants between MSAs and 81,408 pairs of MSAs with nonzero flows, which represent about 70% of all migrations in the original county-to-county Census 2000 data in the U.S. (Table 4.6).

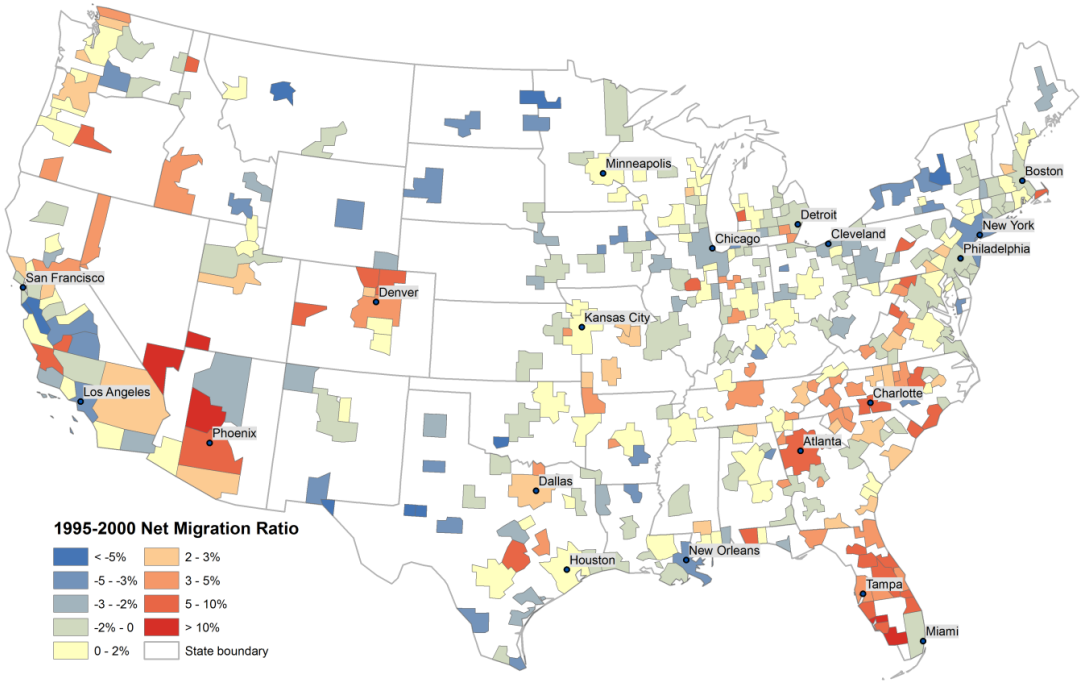


Figure 4.3: Net migration ratio for 358 MSAs with Census 2000 migration data.

Figure 4.4 shows the distribution of competing destination values of MSAs for the entire population (calculatd with Equations 4.7 and 4.8). The northeast has the strongest competing destination while the mid-west has the weakest values. I also used age groups to stratify the MSA migration data. Based on different migration behaviors of the original census age groups (Fotheringham et al. 2004), I regroup migrants into the following seven age groups:

- 5-14 years: preschool/school age;
- 15-19 years: leaving home for university or work;
- 20-24 years: leaving home for university or leaving university for work;
- 25-29 years: leaving university for work, forming couples and starting a family;
- 30-44 years: raising a family;
- 45-59 years: older working age;
- >= 60 years: approaching and beyond retirement age.

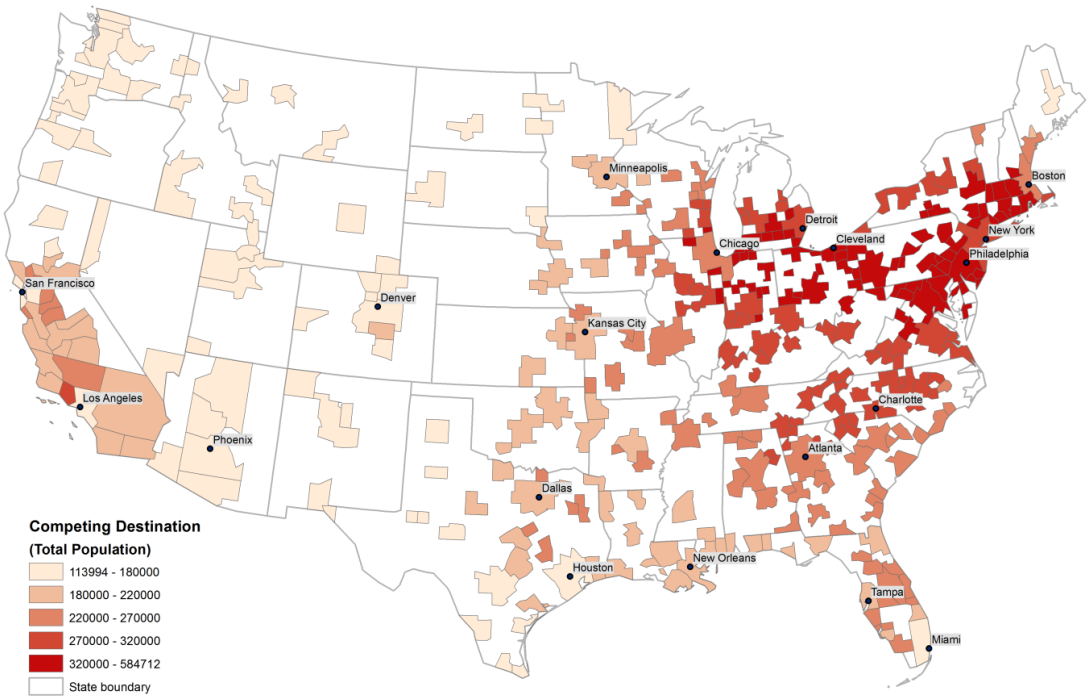


Figure 4.4: Competing destination of 358 MSAs for entire population.

The migration patterns in different age groups can be different in terms of the influence factors and various degree of moving activities. For example, the migration choice of people between age 25 and 29 might be mainly influenced by employment or housing preference, while the choice of older migrants might depend on living cost, local amenities and welfare. The migration rate for the age group 25-29 is 32.4%, which is almost twice of the average migration rate (16.8%) for all age groups.

4.5 Spatial Interaction Modeling of Migration

Through exploratory analysis, I found that migration decay rapidly as distance increases to a certain value and then the rate of change slows down. Therefore a piecewise model is adopted to model the migration characteristics for each of the eight groups (i.e. all-age and seven age groups) within 358 MSAs. The piecewise regression model has a distance breakpoint, beyond which the distance-decay effect is dramatically slowed down. In other words, the overall model consists of two sub-models: one for short-distance migration and the other for long-distance migration. Given that migration flows are count data, the models are estimated with the Poisson regression. The distance break point is determined by minimizing the sum of log-likelihood in Poisson regression for the overall model. Table 4.7 shows the model configuration results for seven population groups and the entire population. The breakpoint distance values generally increase with age (Figure 4.5).

The estimated coefficients for the four factors are approximately the same among the seven groups (Figure 4.6 and Figure 4.7). The origin population has slightly more positive influence than the destination population does on migration flows for all age

groups except for group 25-29, which means that smaller MSAs (in terms of population) attract proportionally more in-migration (except for age 25-29). This might be due to the fact that larger cities tend to have more job opportunities and thus relatively more attractive to migrants in the age group 25-29, while smaller cities offer a better balance of living cost and quality for families and seniors. The *CD* factor has a much stronger negative impact for short-distance migration than that for long-distance migration. For both short- and long-distance migrants, *CD* has relatively less impact on age groups 15-19, 20-24, and 25-29.

Distance has much stronger influences on short-distance migrants than long-distance migrants. Meanwhile, it is interesting that for long-distance migration distance has a much stronger negative impact on the above-60 age group than on any other group (see Figure 4.7). On the other hand, Figure 4.5 shows that the above-60 age group has the longest distance breakpoint (with which we define short-distance and long-distance migration) and the weakest distance decay parameter for short-distance migration, meaning that senior migrants consider a larger *local* neighborhood for migration choice.

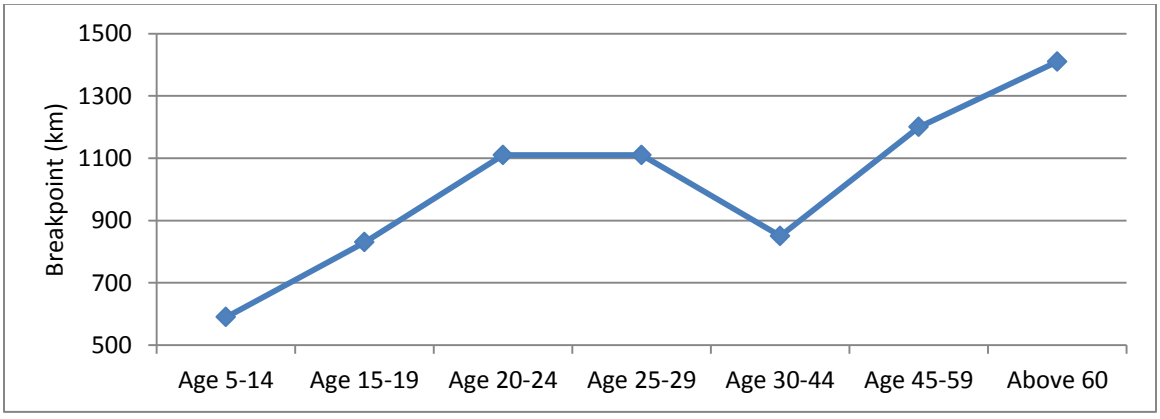


Figure 4.5: Distance breakpoint for each population group, determined by maximizing the Log-likelihood value in Piecewise Poisson regression.

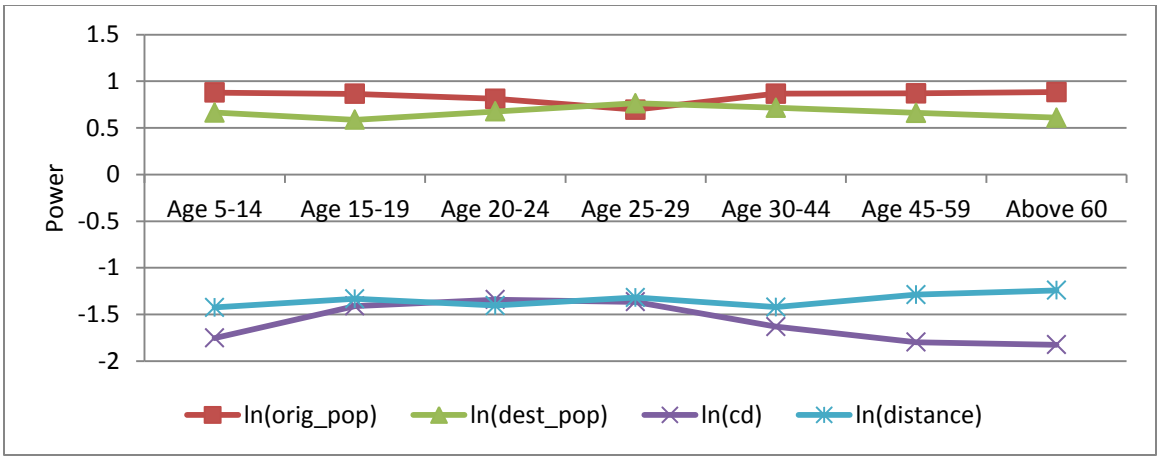


Figure 4.6: Parameter values for *short-distance* migration for each age group.

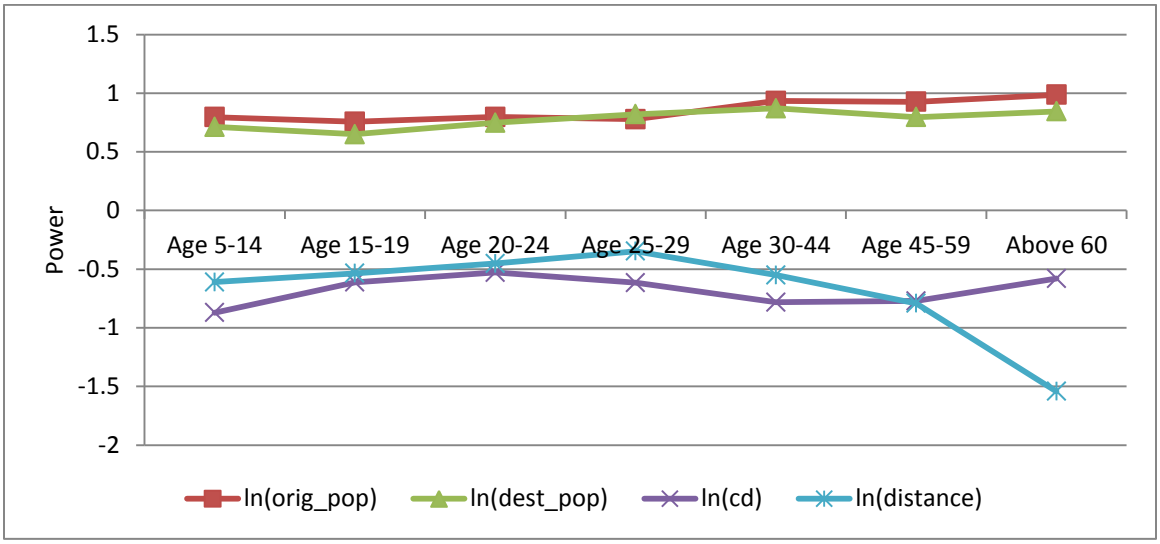


Figure 4.7: Parameters of *long-distance* migration for each age group

Table 4.6: Amount of migrants between MSAs and rural areas by Age Groups.

	Total Pop	5 to 14	15 to 19	20 to 24	25 to 29	30 to 44	45 to 59	60 and over
#Pairs of flow	81409	43564	34973	43402	45388	58196	43514	32912
#Total flow	46629023	6834519	3472409	6145264	6196791	13528458	6154759	4296684
	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
#Flow btw MSA	22966934	3232374	1733195	3193936	3266641	6690144	2899956	1949930
	(49.25%)	(47.29%)	(49.91%)	(51.97%)	(52.72%)	(49.45%)	(47.12%)	(45.38%)
#Flow within MSA	10101023	1569298	580475	931005	1291696	3464804	1417956	845789
	(21.66%)	(22.96%)	(16.72%)	(15.15%)	(20.84%)	(25.61%)	(23.04%)	(19.68%)
#Flow btw MSA and Rural	3041685	500909	270774	433175	345089	754556	400675	336826
	(6.52%)	(7.33%)	(7.8%)	(7.05%)	(5.57%)	(5.58%)	(6.51%)	(7.84%)
#Flow w/in Rural	10519381	1531938	887965	1587148	1293365	2618954	1436172	1164139
	(22.56%)	(22.41%)	(25.57%)	(25.83%)	(20.87%)	(19.36%)	(23.33%)	(27.09%)

Table 4.7: Calibration of migration models for seven population groups and the entire population.

	Total Pop		Age 5-14		Age 15-19		Age 20-24		Age 25-29		Age 30-44		Age 45-59		Above 60	
Breakpoint	1200km		590km		830km		1110km		1110km		850km		1200km		1410km	
	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>
Intercept	14.45 (0.01)	-7.29 (0.02)	13.54 (0.03)	-0.30 (0.03)	10.11 (0.03)	-2.43 (0.05)	10.07 (0.02)	-5.00 (0.05)	10.08 (0.03)	-5.39 (0.04)	12.54 (0.02)	-5.20 (0.03)	13.38 (0.03)	-2.81 (0.04)	13.45 (0.03)	-0.39 (0.06)
ln(orig_pop)	0.89 (2e-4)	1.00 (3e-4)	0.88 (6e-4)	0.80 (6e-4)	0.86 (7e-4)	0.76 (1e-3)	0.81 (5e-4)	0.80 (8e-4)	0.70 (5e-4)	0.78 (7e-4)	0.87 (4e-4)	0.93 (4e-4)	0.87 (5e-4)	0.93 (8e-4)	0.88 (7e-4)	0.99 (1e-3)
ln(dest_pop)	0.73 (2e-4)	0.92 (3e-4)	0.66 (5e-4)	0.72 (6e-4)	0.59 (7e-4)	0.65 (1e-3)	0.68 (5e-4)	0.75 (9e-4)	0.76 (5e-4)	0.82 (7e-7)	0.72 (3e-4)	0.87 (4e-4)	0.66 (5e-4)	0.80 (8e-4)	0.61 (6e-4)	0.84 (1e-3)
ln(cd)	-1.72 (8e-4)	-0.67 (1e-3)	-1.75 (3e-3)	-0.87 (3e-3)	-1.41 (3e-3)	-0.61 (5e-3)	-1.34 (2e-3)	-0.53 (4e-3)	-1.36 (2e-3)	-0.67 (3e-3)	-1.63 (2e-3)	-0.78 (2e-3)	-1.80 (2e-3)	-0.77 (3e-3)	-1.82 (3e-3)	-0.58 (4e-4)
ln(distance)	-1.41 (3e-4)	-0.72 (1e-3)	-1.42 (1e-3)	-0.61 (1e-3)	-1.33 (1e-1)	-0.54 (3e-3)	-1.40 (8e-4)	-0.45 (3e-3)	-1.32 (8e-4)	-0.35 (3e-3)	-1.42 (7e-4)	-0.55 (1e-3)	-1.29 (9e-4)	-0.79 (3e-3)	-1.24 (1e-3)	-1.54 (4e-3)
Log-likelihood	-7023673		-1198627		-713876		-1323255		-1196142		-2121484		-1026685		-955529	
PM	29.3727		31.2186		32.6013		33.0375		31.0709		29.1435		30.6137		35.8861	
SRMSE	4.1282		4.0798		2.7876		2.8778		2.8336		3.6482		3.3692		4.7561	
PSI	0.5605		0.5921		0.617		0.6241		0.5898		0.556		0.5823		0.6717	

Note: All the breakpoints and coefficients are significant at the 0.001 level.

4.6 RESIDUAL MAPPING AND EXPLORATORY ANALYSIS

While important and interesting, the above trends are only the global patterns captured by the model, which cannot reveal local deviations. Moreover, the models can only explain part of the data variation. It is therefore of importance to further analyze the model residuals, discover local variations, and obtain a comprehensive understanding of migration patterns.

4.6.1 MODEL RESIDUAL MAP

A *flow residual* is the difference between the actual flow and its model predicted flow. The *net-residual* for a pair of locations A and B is the absolute difference between the flow *residuals* in two directions, i.e., AB and BA. The direction of a net-residual is the same as the flow with the larger residual. As in Poisson regression, residual is represented by *Standardized Deviance Residual (SDR)*, which is defined in Equation 4.14. *SDR* is a measure of how the regression model fits at each observation, which removes the effect of absolute migration volumes and can be used to compare the relative residuals for different migration pairs. Figure 4.8 shows the top 1000 net-residual (out of 81408, one for each pair of MSAs), with the net migration ratio map in the background.

The design of the residual flow map (Figure 4.8) is as follows. The color of each flow line starts at the origin in green, fades out gradually and disappears at the one-third length of the flow, then emerges in light red at the two-third of the flow length, and becomes more saturated in red when approaching its destination. The purposes of this design are (1) to reduce the overlapping of flow lines and enhance the clarity of the flow map by making the middle part of each flow line transparent; and (2) to better distinguish

origins and destinations by showing them in two different color hues. The flow line width represents the flow (residual) value. While migration trends from the north to the south have been reported in the literature (e.g., Hunt et al. 2008), the residual map in Figure 4.8 and the analyses presented in the remainder of this Chapter present a much more comprehensive view of overall migration patterns and detailed local variations for different age groups, which cannot be extracted with existing methodologies.

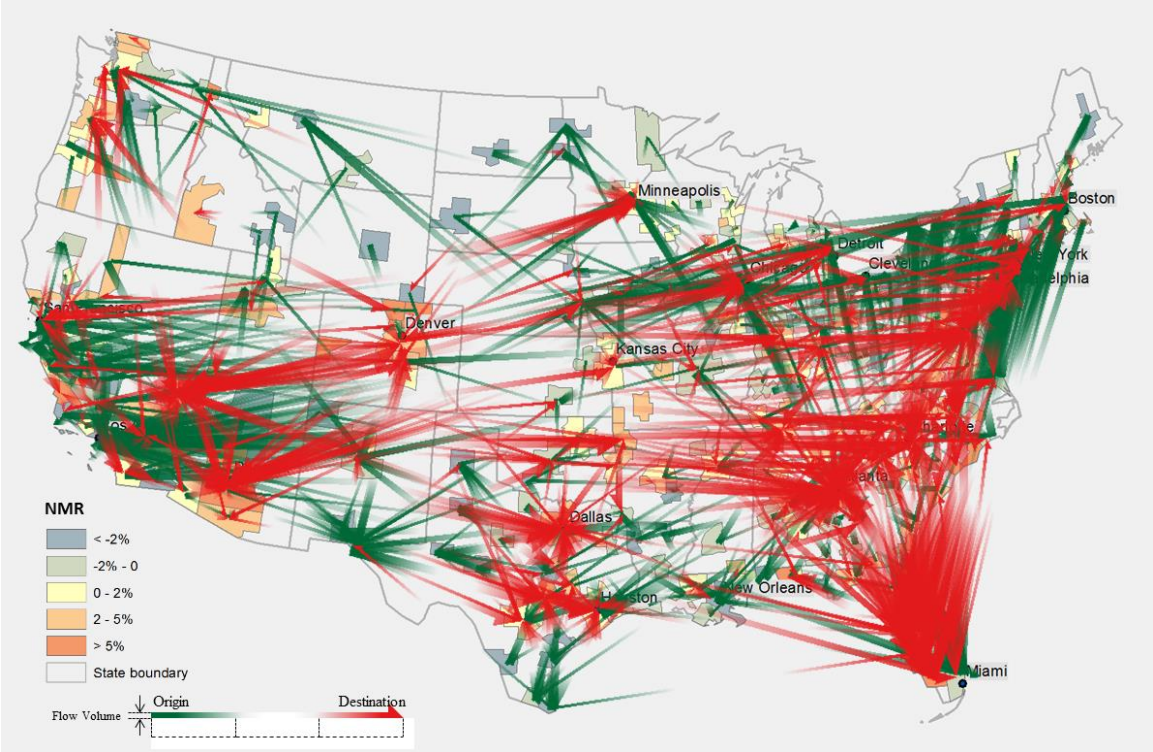


Figure 4.8: Top 1000 net residuals for entire population, overlaid by *NMR*.

The top 1000 flow residuals are selected because they represent the strongest net-residuals between their origins and destinations. However, the selection is subjective and descriptive without statistical inference. To discern clusters from random variations, I extend the local Moran's *I* statistic to assess the significance of local spatial clustering of

residuals. I will examine the statistical patterns in next section with local spatial clustering measures.

4.6.2 Local Association of Residuals

The traditional local Moran's *I* (Anselin 1995) is extended to measure local association of flow residuals by defining a spatial neighborhood for a flow (x_{ij}) from origin *i* to destination *j*, which is the set of flows that starts from the neighborhood J_i of *i* and ends at the neighborhood J_j of *j*. The first-order rook contiguity is used to define J_i and J_j. Figure 4.9 shows an illustrative example, where each neighborhood has five spatial units (including *i* and *j*) and all flows (except x_{ij}) from J_i to J_j are considered neighbors of flow x_{ij}. Here it assumes that origin *i* to destination *j* are not neighbors to each other. In other words, the local Moran's *I* measure for flows between neighbors is not calculated. This neighborhood definition for a flow is a combination of the four definitions proposed in (Chun et al. 2012).

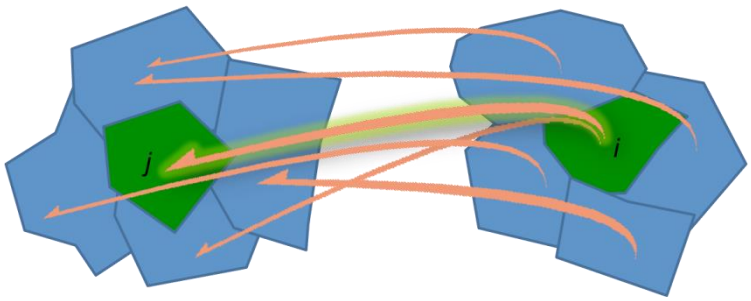


Figure 4.9: Flow neighborhood. All flows except x_{ij} shown in the map are neighbors to flow x_{ij}.

Following the conventional calculation of local Moran's *I* for lattice data (Anselin 1995), I define the local Moran's *I* statistic for flow x_{ij} as:

$$I_{ij} = \frac{x_{ij} - \bar{x}}{\frac{\sum_{k=1}^{n}\sum_{h=1}^{n}(x_{kh} - \bar{x})^2}{N-1}} \left(\left(\sum_{k=1}^{n}\sum_{h=1}^{n} w_{ik} w_{jh} (x_{kh} - \bar{x}) \right) - \left(x_{ij} - \bar{x} \right) \right) \quad (4.19)$$

where N is the total number of flows in the data, w_{ij} indicates the spatial contiguity between i and j (w_{ij} =1 if contiguous, otherwise 0, w_{ii} = 1), and $\bar{x}$ is the mean of all N flows. The measure essentially has three components: the focal flow's deviation from the mean ($x_{ij} - \bar{x}$), neighboring flows' total deviation (excluding the focal flow), and a constant factor ($\frac{\sum_{k=1}^{n}\sum_{h=1}^{n}(x_{kh}-\bar{x})^2}{N-1}$). If flow values represent residuals, $\bar{x}$ = 0 and Equation 4.19 can be simplified as:

$$I_{ij} = \frac{x_{ij}}{\frac{\sum_{k=1}^{n}\sum_{h=1}^{n}(x_{kh})^2}{N-1}} \left(\sum_{k=1}^{n}\sum_{h=1}^{n} w_{ik} w_{jh}\, x_{kh} - x_{ij} \right) \quad (4.20)$$

The significance test of a local Moran's *I* value in this research is based on its *z*-score, as introduced in (Anselin 1995), except for using different spatial objects (i.e., flows instead of area units) and a different neighborhood definition as defined above. To address the multiple-testing problem, the simple Bonferroni adjustment is applied, i.e., modify the significance level α with α/N. A positive z-score indicates the feature is surrounded by features with similar values (either high or low). Here I only focus on high-high spatial association of flow residuals.

Figure 4.10 shows the flow net residuals for the entire population with significant high-high spatial association (Moran's *I p-value* < 0.05). Net migration ratio map is overlaid to verify the results. Based on the direction and the length of the visible portion of a flow in map, it is apparent to infer where the origin and destination are. Most migrants in the Northeast moved to Florida. North and South Carolina and Atlanta area received high in-migration originated from California and Northeast. Arizona is a hot

destination for migrants from Washington and Midwest states including Illinois, Minnesota, Wisconsin, Iowa, North and South Dakota. Las Vegas and Denver are also very hot for migrants from the West.

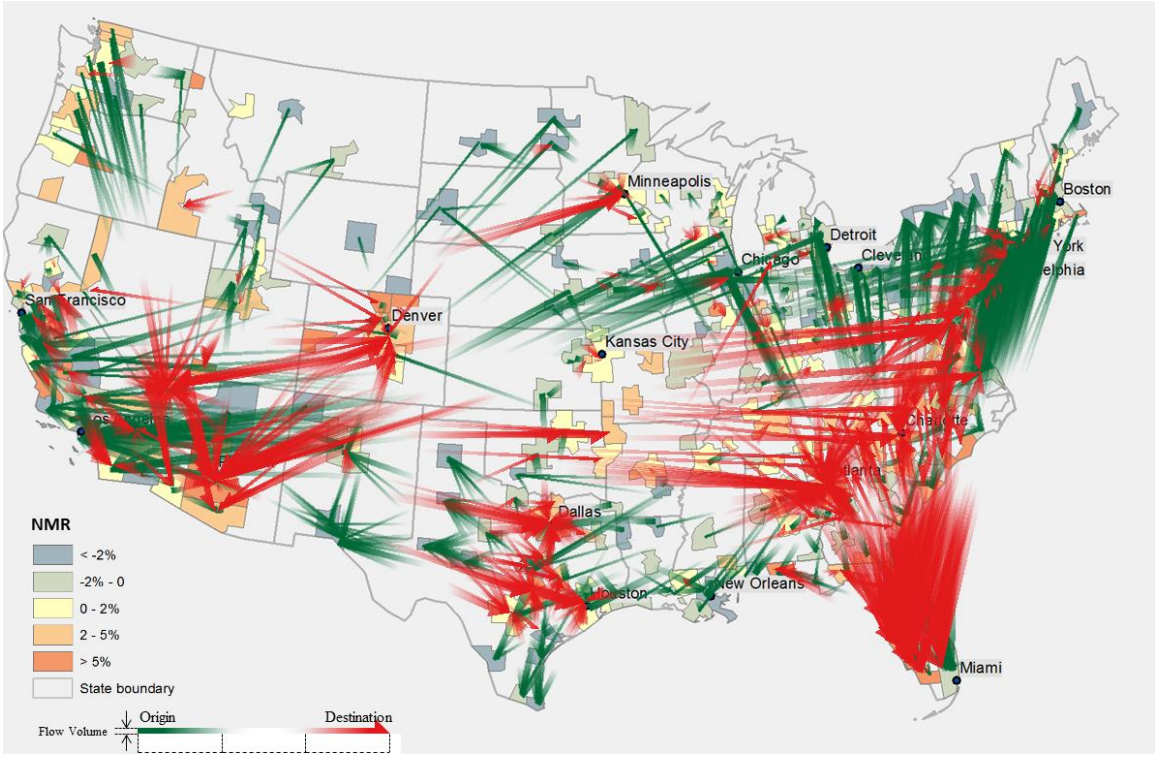


Figure 4.10: Spatial autocorrelation of net residuals for all migration, overlaid by *NMR*.

Figure 4.11 to Figure 4.17 show spatial autocorrelation of net residuals for seven age groups, and net migration ratio map for each age group are overlaid, respectively. Age groups 5-14 (Figure 4.11) and 30-44 (Figure 4.15) show similar pattern since children in age group 5-14 are always moving with their parents (age group 30-44) other than moving independently. The maps suggest that there is a strong trend leaving from large cities. People in those age groups start to raise a family and their migration

intentions are commonly driven by housing preference and living cost. The maps give evidences and comprehensively explain where they are moving to.

Age groups 15-19 (Figure 4.12) and 20-24 (Figure 4.13) show similar pattern, the reason of which might be that people in both groups are leaving home for universities or leaving universities for work. The maps indicate that most of migrations are moving to nearby cities.

Very strong migration patterns to large cities are observed for Age group 25-29 (Figure 4.14) because they are majorly driven by job opportunities. Hot destinations for this age group during 1995-2000 include the San Francisco Bay area, Boston area, Florida, Texas, Atlanta, North Carolina, and Denver.

Two older population groups, age group 45-59 (Figure 4.13) and above 60 (Figure 4.14), demonstrate similar patterns, although the latter (age group above 60) has slightly stronger patterns. Two dramatically hot destinations are detected, which are Arizona and Florida, as expected. For the maps, the origins of those migrants are drawn clearly. Migrants to Florida are mostly from Northeast and Midwest, while migrants to Arizona are from Northwest and Midwest.

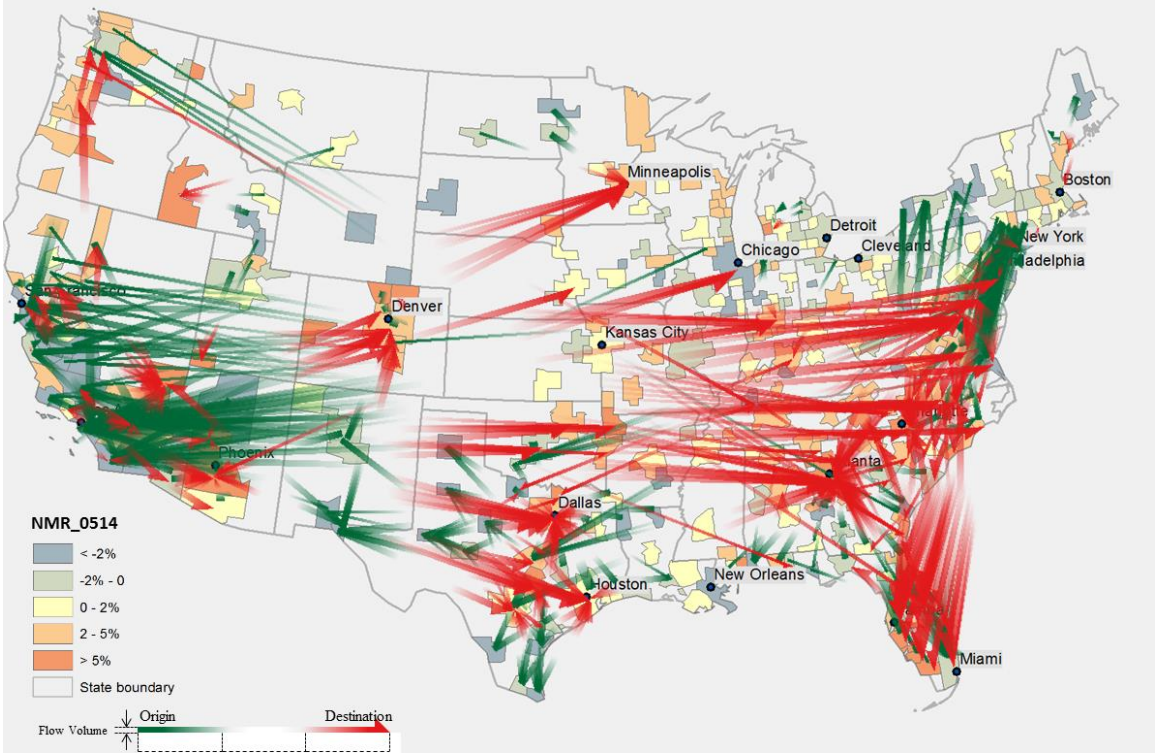


Figure 4.11: Spatial autocorrelation of net residuals for age group 5-14, overlaid by *NMR*.

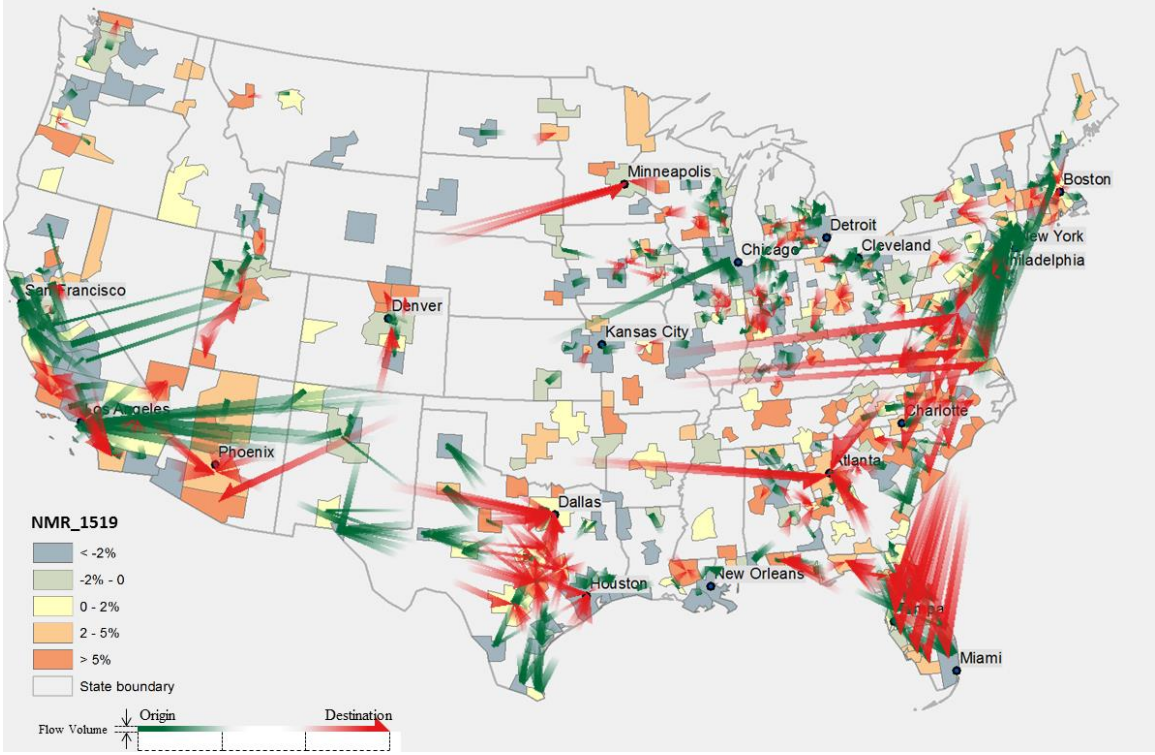


Figure 4.12: Spatial autocorrelation of net residuals of age group 15-19.

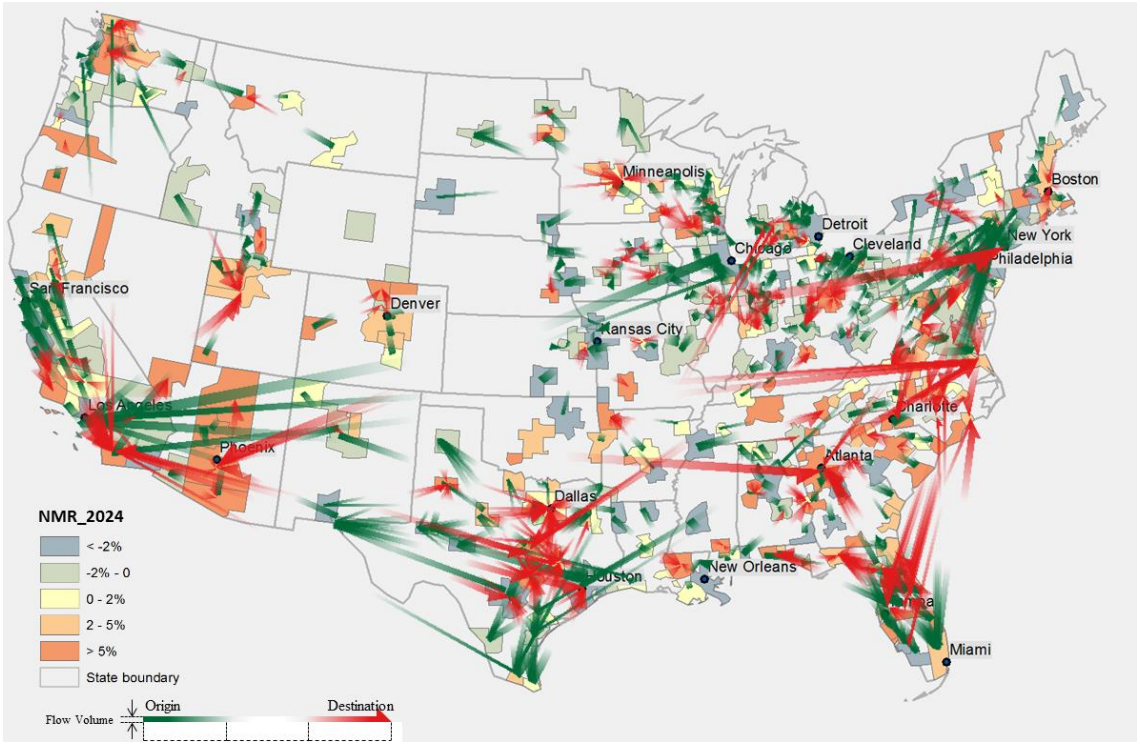


Figure 4.13: Spatial autocorrelation of net residuals for age group 20-24, overlaid by *NMR*.

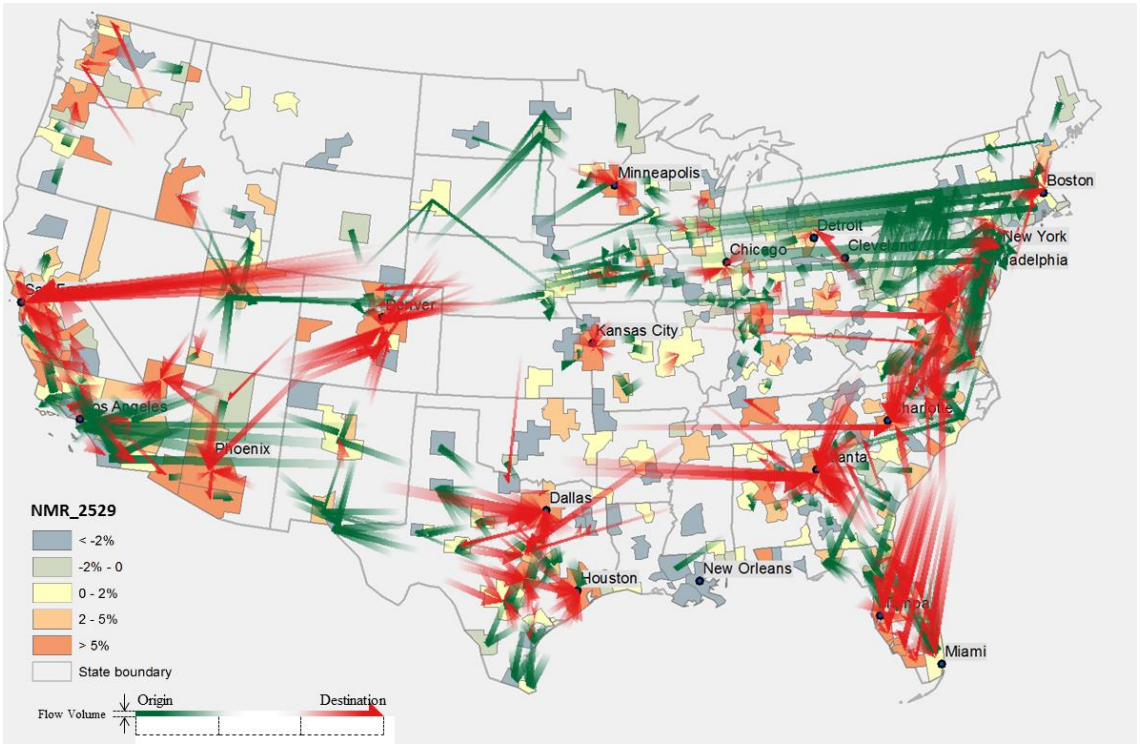


Figure 4.14: Spatial autocorrelation of net residuals for age group 25-29.

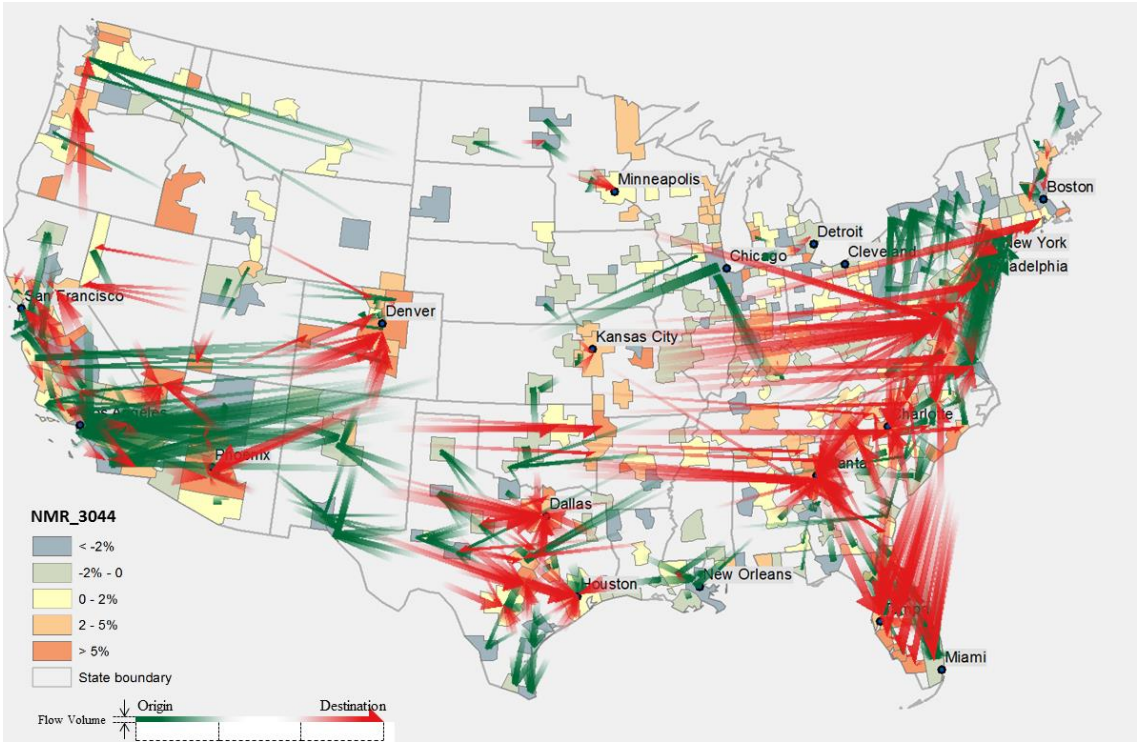


Figure 4.15: Spatial autocorrelation of net residuals for age group 30-44.

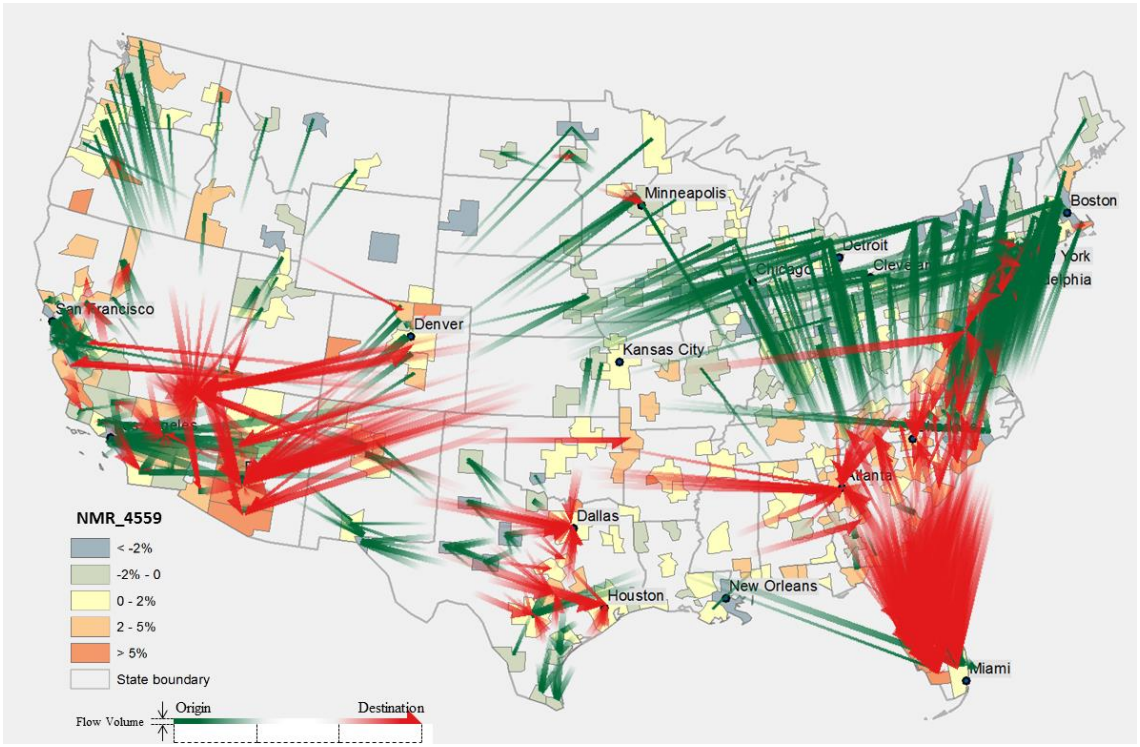


Figure 4.16: Spatial autocorrelation of net residuals for age group 45-59.

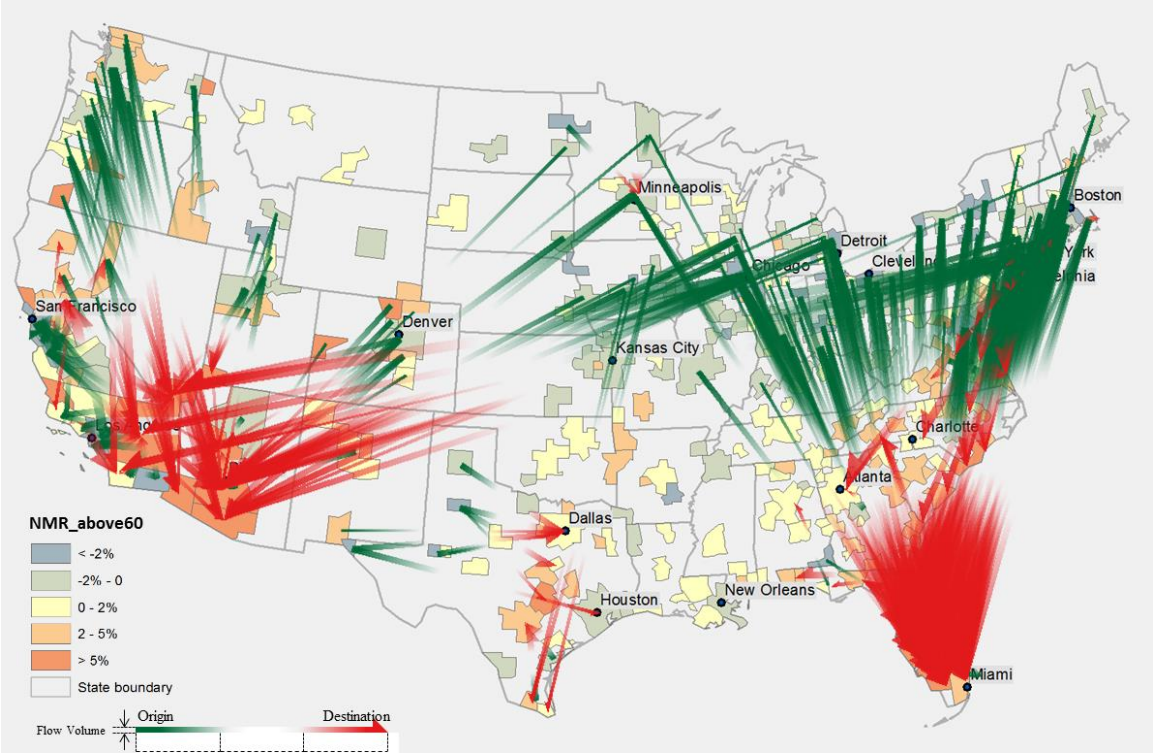


Figure 4.17: Spatial autocorrelation of net residuals for age group above 60.

4.7 Discussion and Conclusion

In this chapter I presented a framework for the extraction of migration patterns with both model configuration and exploratory analysis of model residuals. Different from existing researches that primarily focus on model calibration and global pattern, my approach focuses more on model residuals and local patterns such as spatial clustering and spatial autocorrelation, which are usually not detectable by a global model alone. The spatial autocorrelation results point to significant flow residuals mediated by other socio-cultural and economic factors besides size and distance, accessibility variable considered in modeling stage. I analyzed MSA-to-MSA migration data set in the U.S. for Census 2000 and present a series of patterns for seven age groups discovered from the data,

including both global patterns captured by the models and local patterns deviated from the global trend, which evidence effective of my approach.

To measure the local spatial autocorrelation for each pair of spatial interactions, Local Moran's *I* for spatial interaction data is proposed. For each pair, it assesses the extent of significant spatial clustering of similar values around the pair. In the evaluation of this work, the Local Moran's *I* for spatial interaction works impressively to identify the hot migration flows which are surrounded by similar hot flows. It could be also used to identify flows surrounded by flows with dissimilar values by investigating the negative values.

I used MSAs for analysis instead of using counties or states. Although not presented in this paper, I have tried the approach with counties and states as the aggregation units. County-level results are very unstable with large residuals and variance due to the small area problem. State-level results are more stable than county-level models but cannot detect interesting local patterns due to the coarse spatial resolution and unbalanced unit sizes.

In this research the gravity model considers distance, mass and competing destination variable, and residual analysis captures the other factors manipulating the migration behaviors. Future research could apply various factors in the model and examine patterns of hidden factors by investigating the residual distribution. What's more, although this framework is developed and applied in the context of human migration, it may also be used in other spatial data applications, such as economic activities, trade analysis, animal migration, disease outbreaks, and so on.

CHAPTER 5 : CONCLUSION

5.1 SUMMARY OF RESULTS

This dissertation work develops a series of methodologies to investigate and extract significant patterns in spatial and spatial interaction data. The presented methods have been demonstrated and evaluated with case studies involving both real-world data and synthetic data, featuring both the advancement of methodologies and practical applications.

Chapter 2 presents two new spatial scan statistics with a simple and a hierarchical merge strategy in order to improve existing methods by incorporating smoothing and regionalization techniques. The objective of the new spatial scan statistics is to 1) discover clusters with irregular shapes, 2) alleviate the small-area problem, and 3) alleviate the multiple-testing problem. Synthetic benchmark data sets with circular and irregular clusters are used to evaluate the performance of the proposed methods in terms of statistical power and accuracy measures, including *sensitivity*, *positive predictive value* and *misclassification rate*. Evaluation results suggest that the simple merge method has a comparable power but unbalanced performance on *sensitive* and *ppv*; the hierarchical merge method has both good power and balanced performance on accuracy measures. Robustness analyses indicate that the new methods are not sensitive to different smoothing models.

Chapter 3 presents a new flow scan statistic for spatial interaction data, which is designed to uncover the significant flow patterns. Instead of using a single scanning window as in existing spatial scan statistics, the new method applies a flow tube, which consists of a circular window on the origin and a circular window on the destination, to scan spatial interaction data and discover flow clusters. A statistical measure based on *GLR*, which is independent from neighbourhood size (e.g., population at the origin and destination), is developed as the test statistic for flow scanning. Monte Carlo simulation is adopted to generate a null distribution of *GLR* to enable significance testing of flow clusters. Evaluations with case studies using both area-based and point-based spatial interaction data have demonstrated the detection power and effectiveness of the new flow scan statistic.

Chapter 4 introduces an exploratory framework for the analysis of global and local patterns in spatial interaction data. The framework consists of three components: 1) a gravity model to discover global patterns, taking into consideration factors including distance, mass and competing destination variables; 2) an extended local Moran's *I* to discover spatial clustering of residuals in the flow model, which enables the detection of local patterns; and 3) a novel flow mapping technique to visualize local flow patterns for visual interpretation and understanding. To evaluate the framework, the U.S. internal migration data among 358 Metropolitan Statistical Areas in Census 2000 is stratified into seven age groups and analyzed by applying this newly designed framework. Interesting migration patterns are discovered for each age group, which existing methods cannot detect and compare. The results show that migration patterns in each age group are different but to some degree related. Migrants in age groups 15-19 and 20-24 tend to

move to nearby cities for education. Large cities with more job opportunities are more attractive to people in the age group 25-29. Patterns of movers in the age group 05-14 are closely correlated with those of the age group 30-44, because children always move with their parents. Florida and Arizona are considerably hot destinations for migrants in the age groups of 45-59 and above 60.

The flow scan statistic (Chapter 2) and local flow Moran's *I* statistic (Chapter 3) are different in several aspects. First, the flow scan statistic detects significant spatial flow clusters with more-than-expected flows, while local flow Moran's *I* measures spatial autocorrelation in spatial flows. In addition to the spatial association of high-high or low-low values, the local Moran's *I* could also discover large flows surrounded by low flows or low flow surrounded by high flows. Second, the local Moran's *I* for spatial interaction data could be used to assess patterns of net flows, while the proposed flow scan statistic is not able to handle net flows because the *GLR* is not meaningful for net flows. Third, the flow scan statistic is able to deal with point-based spatial interaction, while Local Moran's *I* could not measure the association of point-based data because each flow in point-based data only represents one individual movement (and hence they are all equal in value).

The overall goal of this dissertation work is to develop models, algorithms and frameworks for extracting statistically significant patterns from spatial lattice and spatial interaction data. The proposed methodologies can potentially be extended to analyze temporal trends or spatio-temporal patterns in spatial lattice or spatial interaction data.

5.2 Limitations and Future Directions

The developed approaches, in their current form, also have several limitations. Although the new spatial scan statistics is designed to detect clusters with different shapes, due to incorporation of smoothing techniques, the proposed spatial scan statistics, theoretically, are not capable of capturing individual clusters (only one unit in the cluster). They assume one single outstanding unit as the random noise and thus exclude its possibility of being a cluster.

The construction strategy of flow tubes used in the flow scan statistic could be further improved. Currently it uses a circular base on each end of the tube, which could be replaced by more comprehensive approach to detect flow clusters between irregular-shaped regions. This is similar to the situation for traditional spatial scan statics. The idea of spatial scan statistic with smoothing and regionalization methods could be borrowed. The challenge is that, a more complicated search strategy with irregular-shaped bases would dramatically increase the computational cost.

The presented approaches for spatial interaction analysis (Chapter 3 and Chapter 4) do not consider the temporal dimension, for which future work is needed since the time dimension is inherent in SI data. The series of methodologies and framework introduced in this dissertation can be extended to capture spatio-temporal patterns in spatial interaction data.

From the implementation perspective, the presented methods of scan statistics using Monte Carlo simulation could take advantage of parallel computing because each simulation is independent of others. An implementation with parallel computing capabilities would reduce the computing time, which can be useful in practice.

REFERENCES

AAMODT G, SAMUELSEN S & SKRONDAL A (2006) A simulation study of three methods for detecting disease clusters. International Journal of Health Geographics **5,** 15.

AMBINAKUDIGE S & PARISI D (2010) Internal migration effectiveness and income effectiveness in the most populous cities in the United States. Population Review **49**.

ANDERSSON C, FRENKEN K & HCLLERVIK A (2006) A complex network approach to urban growth. Environment and Planning A **38,** 1941-1964.

ANSELIN L (1995) LOCAL INDICATORS OF SPATIAL ASSOCIATION - LISA. Geographical Analysis **27**.

ASSUNCAO R, COSTA M, TAVARES A & FERREIRA S (2006) Fast detection of arbitrarily shaped disease clusters. Statistics in Medicine **25,** 723-742.

AYENI B & REFEREE J.B.H. RAMSEY (1983) Algorithm 11: Information statistics for comparing predicted and observed trip matrices. Environment and Planning A **15,** 1259-1266.

BALCAN D, COLIZZA V, GONCALVES B, HU H, RAMASCO JJ & VESPIGNANI A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences of the United States of America **106,** 21484-21489.

BARTHÉLEMY M (2011) Spatial networks. Physics Reports **499,** 1-101.

BERGLUND S & KARLSTRÖM A (1999) Identifying local spatial association in flow data. Journal of Geographical Systems **1,** 219-236.

BLACK WR (1992) Network autocorrelation in transport network and flow systems. Geographical Analysis **24,** 207-222.

BLACK WR & THOMAS I (1998) Accidents on belgium's motorways: a network autocorrelation analysis. Journal of Transport Geography **6,** 23-31.

BROCKMANN D, HUFNAGEL L & GEISEL T (2006) The scaling laws of human travel. Nature **439,** 462-465.

CECCATO V (2005) Homicide in Sao Paulo, Brazil: Assessing spatial-temporal and weather variations. Journal of Environmental Psychology **25,** 307-321.

CHAN HP (2009) DETECTION OF SPATIAL CLUSTERING WITH AVERAGE LIKELIHOOD RATIO TEST STATISTICS. Annals of Statistics **37,** 3985-4010.

CHUN Y (2008) Modeling network autocorrelation within migration flows by eigenvector spatial filtering. Journal of Geographical Systems **10,** 317–344.

CHUN Y, KIM H & KIM C (2012) Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: An application of the US interstate commodity flows. Computers, Environment and Urban Systems.

CLARK C (1967) Population growth and land use. Macmillan, London.

COHEN JE, ROIG M, REUMAN DC & GOGWILT C (2008) International migration beyond gravity: A statistical model for use in population projections. PNAS **105** 15269-15274.

COSTA MA, ASSUNCAO RM & KULLDORFF M (2012) Constrained spanning tree algorithms for irregularly-shaped spatial clustering. Computational Statistics & Data Analysis **56,** 1771-1783.

CUI W, ZHOU H, QU H, WONG PC & LI X (2008) Geometry-Based Edge Clustering for Graph Visualization. Ieee Transactions on Visualization and Computer Graphics **14**.

DONGES JF, HEITZIG J, DONNER RV & KURTHS J (2012) Analytical framework for recurrence network analysis of time series. Physical Review E **85**.

DUCZMAL L & ASSUNCAO R (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics & Data Analysis **45,** 269-286.

DUCZMAL L, CANÇADO ALF, TAKAHASHI RHC & BESSEGATO LF (2007) A genetic algorithm for irregularly shaped spatial scan statistics. Computational Statistics & Data Analysis **52,** 43-52.

DUCZMAL L, KULLDORFF M & HUANG L (2006) Evaluation of Spatial Scan Statistics for Irregularly Shaped Clusters. Journal of Computational & Graphical Statistics **15,** 428-442.

ERLANDER S & STEWART NF (1990) The Gravity Model in Transportation Analysis: Theory and Extensions. VSP, Utrecht, The Netherlands.

EUBANK S, GUCLU H, KUMAR VSA, et al. (2004) Modelling disease outbreaks in realistic urban social networks. Nature **429**.

FANG ZX, SHAW SL, TU W, LI QQ & LI YG (2012) Spatiotemporal analysis of critical transportation links based on time geographic concepts: a case study of critical bridges in Wuhan, China. Journal of Transport Geography **23,** 44-59.

FERGUSON NM, CUMMINGS DAT, CAUCHEMEZ S, et al. (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature **437,** 209-214.

FERGUSON NM, CUMMINGS DAT, FRASER C, CAJKA JC, COOLEY PC & BURKE DS (2006) Strategies for mitigating an influenza pandemic. Nature **442**.

FISCHER MM & GOPAL S (1994) Artificial neural networks: a new approach to modeling interregional telecommunication flows. Journal of Regional Science **34,** 503–527.

FISCHER MM & GRIFFITH DA (2008) Modeling spatial autocorrelation in spatial interaction data: an application to patent citation data in the european union. Journal of Regional Science **48**.

FISCHER MM, REISMANN M & HLAVACKOVA-SCHINDLER K (2003) Neural Network Modeling of Constrained Spatial Interaction Flows: Design, Estimation, and Performance Issues. Journal of Regional Science **43,** 35–61.

FORTUNATO S (2010) Community detection in graphs. Physics Reports-Review Section of Physics Letters **486,** 75-174.

FOTHERINGHAM AS (1983) A New Set of Spatial-Interaction Models: The Theory of Competing Destinations. Environment and Planning A **15,** 15-36.

FOTHERINGHAM AS (1986) Further discussion on distance-deterrence parameters and the competing destinations model. Environment and Planning A **18,** 553–556.

FOTHERINGHAM AS, BRUNSDON C & CHARLTON M (2000) Quantitative Geography. SAGE Publications, London/Thousand Oaks/New Delhi.

FOTHERINGHAM AS, REES P, CHAMPION T, KALOGIROU S & TREMAYNE AR (2004) The development of a migration model for England and Wales: overview and modelling out-migration. Environment and Planning A **36,** 1633-1672.

GEARY RC (1954) The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician **5,** 115-146.

GENNADY A, NATALIA A & STEFAN W (2007) Visual analytics tools for analysis of movement data. SIGKDD Explor. Newsl. **9,** 38-46.

GERMANN TC, KADAU K, LONGINI IM & MACKEN CA (2006) Mitigation strategies for pandemic influenza in the United States. Proceedings of the National Academy of Sciences of the United States of America **103**.

GETIS A (2008) A history of the concept of spatial autocorrelation: A geographer's perspective. Geographical Analysis **40,** 297-309.

GETIS A & ORD JK (1992) The analysis of spatial association by use of distance statistics. Geographical Analysis **24,** 189-206.

GOH S, LEE K, PARK JS & CHOI MY (2012) Modification of the gravity model and application to the metropolitan Seoul subway system. Physical Review E **86,** 6.

GULDMANN J-M (1999) Competing destinations and intervening opportunities interaction models of inter-city telecommunication flows. Papers in Regional Science **78,** 179–194.

GUO D (2007) Visual analytics of spatial interaction patterns for pandemic decision support. International Journal of Geographical Information Science **21**.

GUO D (2008) Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science **22,** 801-823.

GUO D (2009) Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data. IEEE Transactions on Visualization and Computer Graphics (TVCG: Proc. of InfoVis'09) **15,** 1041-1048.

GUO D & WANG H (2011) Automatic Region Building for Spatial Analysis. Transactions in Gis **15,** 29-45.

GUO D & ZHU X (2014) Origin-Destination Flow Data Smoothing and Mapping. IEEE Transactions on Visualization and Computer Graphics **99,** 1.

HEFFERNAN R, MOSTASHARI F, DAS D, KULLDORFF M & WEISS D (2004) Syndromic surveillance in public health practice, New York City. Emerging Infectious Diseases **10,** 858-864.

HOLTEN D (2006) Hierarchical Edge Bundles: Visualization of Ajacency Relations in Hierarchical Data. IEEE Transactions on Visualization and Computer Graphics (TVCG: Proc. of InfoVis'06) **12,** 741-748.

HOLTEN D & WIJK JJV (2009) Force-Directed Edge Bundling for Graph Visualization. Computer Graphics Forum (Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization) **28,** 983-990.

HU P & POOLER J (2002) An empirical test of the competing destinations model. Journal of Geographical Systems **4,** 301-323.

HUNT LL, HUNT MO & FALK WW (2008) Who is Headed South? US Migration Trends in Black and White, 1970-2000. Social Forces **87,** 95-119.

JOHNSON KM, VOSS PR, HAMMER RB, FUGUITT GV & MCNIVEN S (2005) Temporal and spatial variation in age-specific net migration in the United States. Demography **42,** 791-812.

JUNG W-S, WANG F & STANLEY HE (2008) Gravity model in the Korean highway. Epl **81**.

KALUZA P, KOELZSCH A, GASTNER MT & BLASIUS B (2010) The complex network of global cargo ship movements. Journal of the Royal Society Interface **7**.

KAREMERA D, OGULEDO VI & DAVIS B (2000) A gravity model analysis of international migration to North America. Applied Economics **32**.

KNUDSEN DC & FOTHERINGHAM AS (1986) Matrix comparison, goodness-of-fit, and spatial interaction modeling. International Regional Science Review **10,** 127-147.

KULLDORFF M (1997) A Spatial Scan Statistic. Communications in Statistics—Theory and Methods **26,** 1481–1496.

KULLDORFF M, HUANG L, PICKLE L & DUCZMAL L (2006) An elliptic spatial scan statistic. Statistics in Medicine **25,** 3929-3943.

KULLDORFF M, TANGO T & PARK PJ (2003) Power comparisons for disease clustering tests. Computational Statistics & Data Analysis **42,** 665-684.

KWAN MP (2000) Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. Transportation Research Part C-Emerging Technologies **8**.

LAMBIOTTE R, BLONDEL VD, DE KERCHOVE C, et al. (2008) Geographical dispersal of mobile communication networks. Physica a-Statistical Mechanics and Its Applications **387**.

LAUBE P, IMFELD S & WEIBEL R (2005) Discovering relative motion patterns in groups of moving point objects. International Journal of Geographical Information Science **19**.

LENORMAND M, HUET S, GARGIULO F & DEFFUANT G (2012) A Universal Model of Commuting Networks. Plos One **7**.

MCCOOL SF & KRUGER LE (2003) Human migration and natural resources: Implications for land managers and challenges for researchers. US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

MCCULLAGH P & NELDER JA (1989) Generalized Linear Models. Chapman & Hall/CRC.

MORAN PAP (1948) The interpretation of statistical maps. Journal of the Royal Statistical Society Series B-Statistical Methodology **10,** 243-251.

NAUS JI (1995) The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association **60,** 532-538.

NEILL DB, MOORE AW & COOPER GF (2006) A Bayesian spatial scan statistic. Advances in neural information processing systems **18,** 1003.

NEWMAN MEJ (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America **103,** 8577-8582.

NJOCK J-C & WESTLUND L (2010) Migration, resource management and global change: Experiences from fishing communities in West and Central Africa. Marine Policy **34**.

OPENSHAW S (1998) Neural network, genetic, and fuzzy logic models of spatial interaction. Environment and Planning A **30,** 1857-1872.

OPENSHAW S, CHARLTON M, WYMER C & CRAFT A (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. International Journal of Geographical Information Science **1,** 335-358.

ORD JK & GETIS A (1995) Local spatial autocorrelation statistics - distributional issues and an application. Geographical Analysis **27,** 286-306.

PATIL GP & TAILLIE C (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environmental and Ecological Statistics **11,** 183-197.

PERRY MJ (2006) Domestic Net Migration in the United States: 2000 to 2004.).

PHAN D, XIAO L, YEH R & HANRAHAN P (2005) Flow map layout. In: Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on) IEEE, 219-224.

PITFIELD DE (1978) Sub-optimality in freight distribution. Transportation Research **12,** 403-409.

RAE A (2009) From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. Computers, Environment and Urban Systems **33,** 161-178.

ROY JR & THILL J-C (2004) Spatial interaction modelling. Papers in Regional Science **83,** 339-361.

SIMINI F, GONZALEZ MC, MARITAN A & BARABASI A-L (2012) A universal model for mobility and migration patterns. Nature **484**.

STOUFFER SA (1960) Intervening Opportunities and Competing Migrants. Journal of Regional Science **2,** 1-26.

TANGO T (2008) A spatial scan statistic with a restricted likelihood ratio. Japanese Journal of Biometrics **29,** 75-95.

THIEMANN C, THEIS F, GRADY D, BRUNE R & BROCKMANN D (2010) The Structure of Borders in a Small World. Plos One **5**.

THORSEN I & GITLESEN JP (1998) Empirical evaluation of alternative model specifications to predict commuting flows. Journal of Regional Science **38,** 273-292.

TOBLER WR (1976) Spatial interaction patterns. Journal of Environmental Systems **6,** 271-301.

TOBLER WR (1981) A model of geographical movement. Geographical Analysis **13,** 1-20.

TOBLER WR (1987) Experiments in migration mapping by computer. American Cartographer **14,** 155-163.

TOBLER WR (2004) Movement Mapping.).

VERBEEK K, BUCHIN K & SPECKMANN B (2011) Flow Map Layout via Spiral Trees. Ieee Transactions on Visualization and Computer Graphics **17**.

VIBOUD C, BJORNSTAD ON, SMITH DL, SIMONSEN L, MILLER MA & GRENFELL BT (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science **312**.

VITALI S & BATTISTON S (2011) Geography versus topology in the European Ownership Network. New Journal of Physics **13**.

WALTHER G (2010) Optimal and fast detection of spatial clusters with scan statistics. Annals of Statistics **38,** 1010-1033.

WILSON AG (1967) Astatistical theory of spatial distribution models. Transportation Research Part A **1,** 253–269.